SMAR

ANOVA

Today's Topic (SMAR-t): ANOVA

- Analysis of Variance (i.e. ANOVA)
  - Independent Measures ANOVA
  - Repeated Measures ANOVA
  - Mixed Factorials
  - Analysis of Covariance (ANCOVA)
    - Using Covariates

Experimental Designs

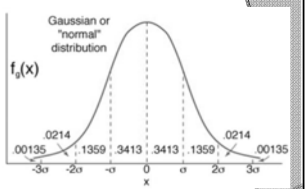
- ANOVA is very common with traditional designs of experiments involving 1 or more "factors," with 2 or more "levels"
  - Factor
    - Level
- Factors can be "between" or "within"
  - A.k.a. Independent/Dependant Measures
  - A.k.a. Grouping/Repeated Factors

Types of Outcomes for ANOVA

- Continuously scaled outcomes assumed to follow the normal distribution, or that can be transformed so that it does (i.e. "normalized")
  - Examples: BMI, BP, BMD, Strength, Standardized Scores, Viral Loads, Force, Averages or Sums of Likert-Scaled items (scale scores), Optical Density, Volume, Response Time, Distance, etc.

Quick Review: Gaussian Distribution Function

- A.K.A. The "Normal Distribution"
- A.K.A. The "Bell-Shaped Curve"
- Has known probabilities associated with it,
- Thus all Parametric Statistics are based on the Gaussian Distribution

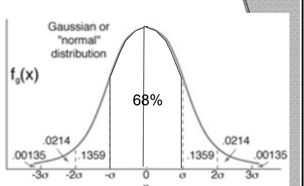


$$f_g(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\bar{x})^2/2\sigma^2}$$

Where $\bar{x}$ = mean, and σ = standard deviation

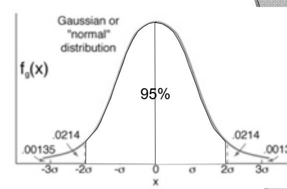
Quick Review: Gaussian Distribution Function

- About 68% of all scores fall within 1 SD unit from the mean.



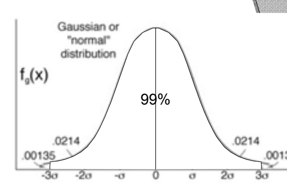
Quick Review: Gaussian Distribution Function

- About 68% of all scores fall within 1 SD unit from the mean.
- About 95% of all scores fall within 2 SD units from the mean.



Quick Review: Gaussian Distribution Function

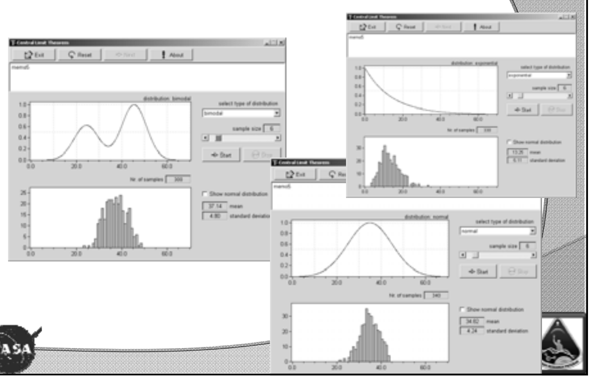
- About 68% of all scores fall within 1 SD unit from the mean.
- About 95% of all scores fall within 2 SD units from the mean.
- About 99% of all scores fall within 3 SD units from the mean.



Central Limit Theorem

- States that for any population with mean μ and standard deviation σ , the distribution of sample means with sample size *n* will approach a *normal distribution* with μ and SD of as n approaches infinity.
- REGARDLESS of the shape of the distribution in the population.
- By the time sample sizes hit around 30, sampling distribution of means is close to normal.

Demo of central limit theorem.

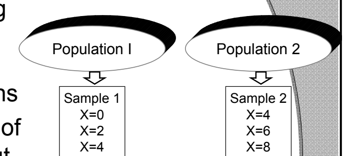


Thus…

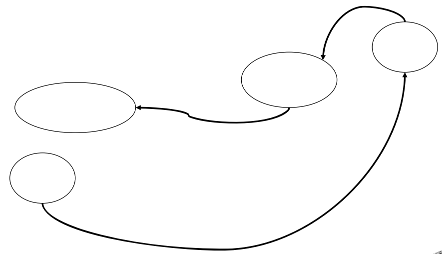
- Since we know so much about the Normal Distribution
- And we know that sample summaries (means or otherwise) tend to follow that distribution
  - Even data collected from non-normal samples
  - Especially so with large sample size (big-n)
- We can usually apply our knowledge of the normal distribution to statistical comparisons, estimates, and probability
  - As long as we do some preliminary screening…

Moving to the t-test for comparing two samples

- Used for comparing two samples collected randomly from two populations
- Many other flavors of the t-test exist… but we'll start here.



Dissect the formula:



Dissect the formula: Numerator

The difference between two sample means

Dissect the formula: Denominator

The difference between two sample means

Divided by some measure of standard error of the differences

Dissect the formula: Question?

The difference between two sample means

Divided by some measure of standard error of the differences

Are the differences that I see between my two means unusual, given variability among other sample means of this size?

T-tests on the Computer:

- Software gives us t-score and a p-value
- Allowing us to test hypotheses that the two samples come from the same population (or not)
- And describe the magnitude of the differences (confidence intervals)
- Ex. $t = 4.87$, $p<.001$
  - H_{null}: Two samples are from same population
  - H_{alt}: Two samples are from different populations
- Reject the Null (alpha < .05) & Report the magnitude of the differences

Virtues of the t-test

- EVERYONE seems to understand it!
- With CLT, it's easy to apply to lots of different data scenarios
- There are other versions that make it very flexible
  - Formula for "Repeated Measures" designs
  - Formula for problems associated with non-normality and/or variance heterogeneity

Hypothesis testing Scenario

- The "null" hypothesis for the t-test is that the two groups come from the same population
  - Thus will have similar means, given sd
- The "alternative" hypothesis is usually that they don't
  - Thus have "different" means, but similar sd
  - Can be directional
- We use the t-statistic in an attempt to Reject the null, supporting our claim of the alternative

Consequences of Hypothesis Testing & Alpha

	The Truth is:	
Your decision is:	**H_0 Really isTrue** (there's no effect)	**H_0 is Actually False** (there is an effect)
You Rejected H_0 Due to a Statistically Significant Result (Conclude the 2 groups must come from different populations)	**Type I Error** Probability = α	**Power** Probability = $(1-\beta)$
You Accepted H_0 Due to a Non-Significant Result (Assume the 2 groups are come from same population)	Probability = $1-\alpha$	**Type II Error** Probability = β

If you have a "significant" result:

Your decision is:	The Truth is: H_0 Really isTrue (there's no effect)	The Truth is: H_0 is Actually False (there is an effect)
You Rejected H_0 Due to a Statistically Significant Result (Conclude the 2 groups must come from different populations)	**Wrong Conclusion**	**Right Conclusion**

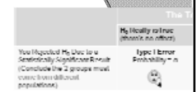
Given a significant t-score comparing means….

If you have a "non-significant" result:

Your decision is:	The Truth is: H_0 Really isTrue (there's no effect)	The Truth is: H_0 is Actually False (there is an effect)
You Accepted H_0 Due to a Non-Significant Result (Assume the 2 groups are come from same population)	**Right Conclusion**	**Wrong Conclusion**

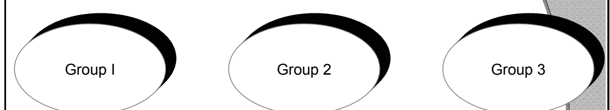
Given a non-significant t-score comparing means….

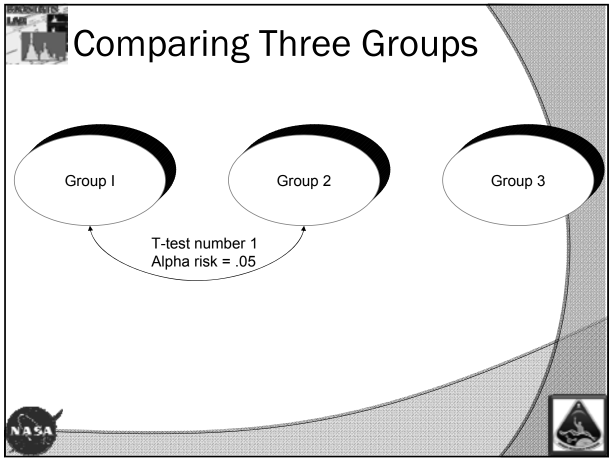
Limitations of t-tests

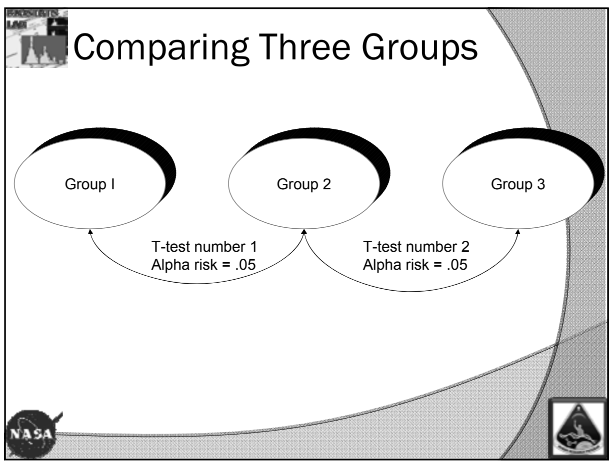


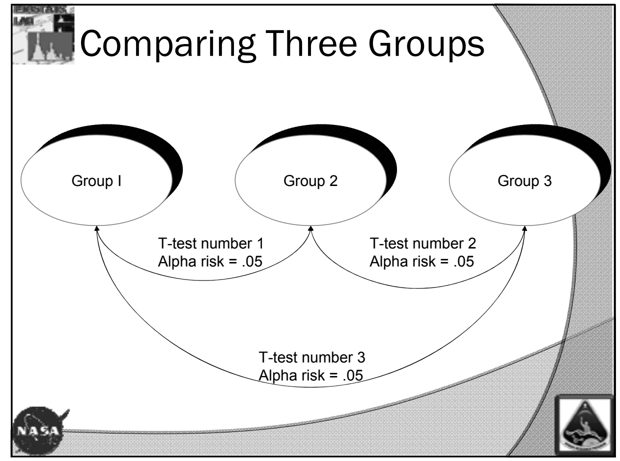
- Alpha risk is .05 for each t-test
  - Probability of falsely rejecting the null, and concluding that there is a difference, when it's really due to chance.
  - So comparing 3, 4, 5 or more groups is quite problematic!

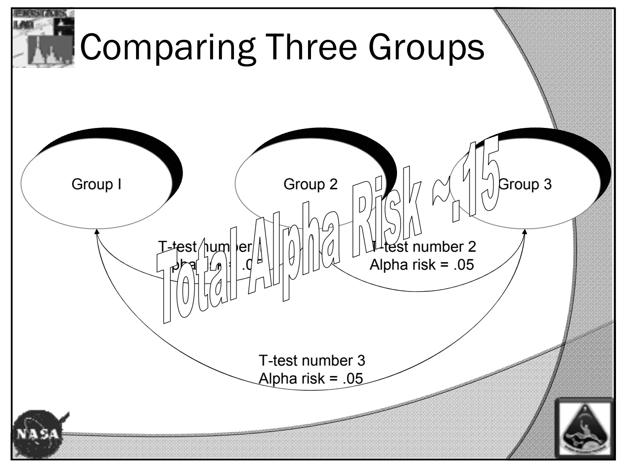
Comparing Three Groups



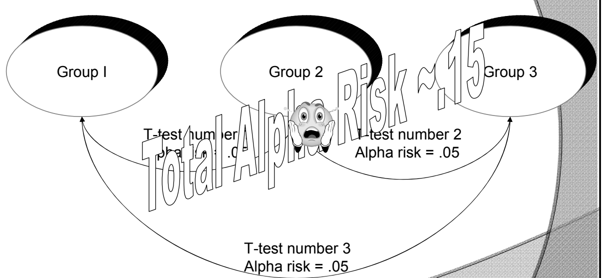

Comparing Three Groups
Group I
Group 2
Group 3
T-test number 1
Alpha risk = .05


Comparing Three Groups
Group I
Group 2
Group 3
T-test number 1
Alpha risk = .05
T-test number 2
Alpha risk = .05


Comparing Three Groups
Group I
Group 2
Group 3
T-test number 1
Alpha risk = .05
T-test number 2
Alpha risk = .05
T-test number 3
Alpha risk = .05


Comparing Three Groups
Group I
Group 2
Group 3
Total Alpha Risk ~.15
T-test number 2
Alpha risk = .05
T-test number 3
Alpha risk = .05

Comparing Three Groups



Analysis of Variance (ANOVA)

- Can compare unlimited number of groups or occurrences, and still keep alpha risk = .05
- Able to take multiple grouping (or time) factors into account and determine their independent and combined effects
- Can examine "trends" in data, and can test specific (often complex) hypotheses
- The analytic focus is on variance, but the interpretation falls back to means—thus results become intuitive

Assumptions Required of ANOVA

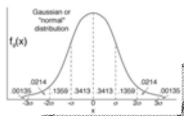
- Data collected randomly from the population, with roughly equal n per cell
  - And sufficiently large n (n>30, common r-o-t)
- Data measured on interval or ratio scale, and is normally distributed
- Homogeneity of variance across groups
- Sphericity for RM designs—variance of the differences between means for any pair of groups is equal to any other pair

Assumption of Randomly Collected Data with Sufficiently Large n

- Is our subject pool at NASA randomly selected from our inference-population?
  - Are those bedrest subjects representative of astronauts?
  - Are today's astronauts representative of future ones?
- Regarding n, How big is big enough?
  - Rule of Thumb… at least 30 per group
  - More is better
    - Cautions about overpowered studies…
  - But BALANCE is critical!!
    - Rule of thumb—smallest group should not be less than 1/3rd the size of the largest group.

Assumption of Interval or Ratio Scale & Normal

- The “bell-shaped” curve—assumption of all parametric statistics
- Studies show that ANOVA is robust to violations of this, but only if sample size is substantially large, and Homogeneity is met

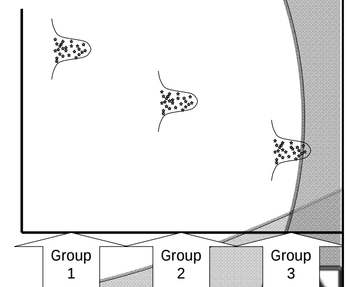


Assumption of Homogeneity of Variance Across Groups

- Variance on the dependant variable should be similar across groups
  - Why?
- Because we’re examining VARIANCE in ANOVA, and so we need for variance in each group to be roughly similar before we can conclude that any differences that we find are attributable to *group* differences (not mere variability differences).
- Even in Means Comparisons (ex.t-tests), since Means are highly affected by variability, we need for variability to be similar in our groups so that differences that we find can be attributed to true group differences, and not merely by variability differences between our groups.

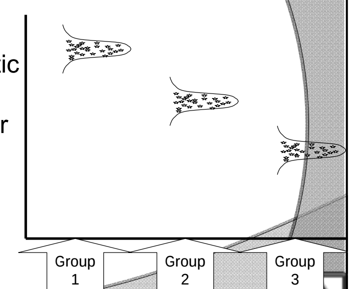
More on Homogeneity of Variance

- If distributions are normal in one, then should be for all



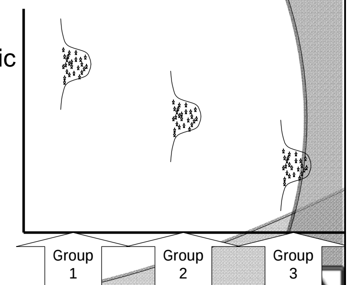
More on Homogeneity of Variance

- If distributions in 1 group is leptokurtotic (tall and skinny), then it should be for all other groups



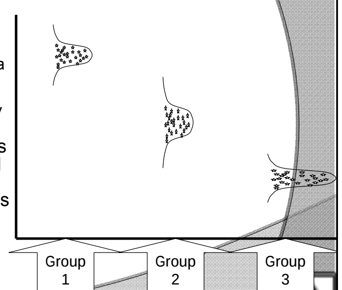
More on Homogeneity of Variance

- If distributions in 1 group is platykurtotic (short & fat) then it should be for all other groups



More on Homogeneity of Variance

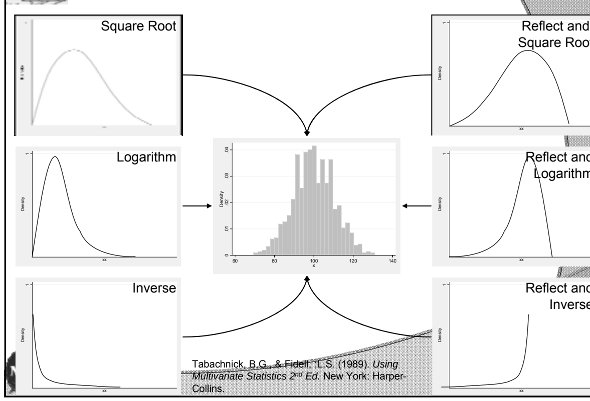
- Any Miss-Match is a Problem
  - Because we might interpret a statistical differences to real group differences, when it's actually due to heterogeneity of variance
- …Thankfully there are ways to test for this problem, and solutions are sometimes possible. SPSS will test this assumption for us (stay tuned)



What about skewed data?

- Positive or negative skews in the data can wreak havoc with statistical analysis
  - Thus always recommend thorough data screening
  - Identify outliers—data entry errors?
  - Consider data transformations if necessary
    - A great thing to google!

Common Transformations



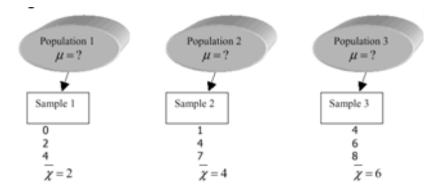
Tabachnick, B.G., & Fidell, L.S. (1989). *Using Multivariate Statistics 2nd Ed.* New York: Harper-Collins.

Two General Types of ANOVA

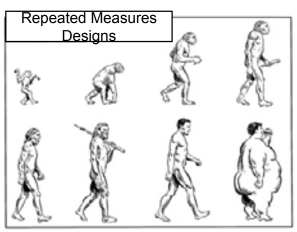
- Independent Measures ANOVA (IM-ANOVA)
  - Data are collected from separate groups of subjects, and comparisons among *groups* are desired
    - Muscle Size by Treatment (controls vs. two intervention groups)
    - Blood Flow by Gender
- Repeated Measures ANOVA (RM-ANOVA)
  - Data are collected from the same group of subjects on multiple occasions/times, and comparisons of *occasions* are desired.
    - Longitudinal Studies
    - Outcomes measured at L-10, L-1, L+2, L+5 L+25, R+2…
    - Balance Scores Pre Bedrest, During and Post Bedrest
- Mixed Factorial
  - Mix of IM and RM factors in the same experiments
    - Gender (m,f) by Time (pre, post) effects

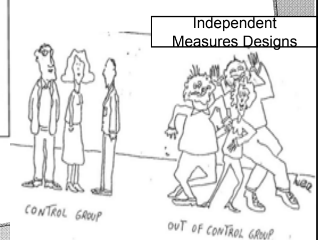
One-Way IM-ANOVA

- For comparing two or more populations
  - Where sample data have been collected

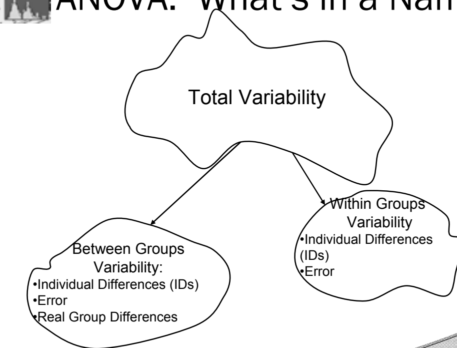


IM & RM Designs…





ANOVA: What's in a Name?



Analysis of Variance F-Ratio

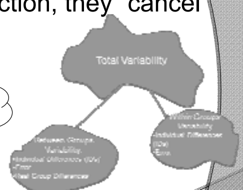
- ANOVA is truly an analysis of a measure of *variability*, called "*variance*," that measures and separates variability attributable to
  - Within-Groups Variability
  - Between-Groups Variability
- We Evaluate an "F-Ratio" Representing the Ratio of B/T over W/I:

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's} + \text{error} + \text{group differences}}{\text{ID's} + \text{error}}$$

Recall your Simple Algebra…

- If the same quantity exists in the Numerator and Denominator of a fraction, they "cancel each other out"

Assuming homogeneity of variance



The F-Ratio

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\cancel{\text{ID's}} + \cancel{\text{error}} + \text{group differences}}{\cancel{\text{ID's}} + \cancel{\text{error}}}$$

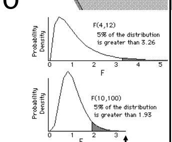
Recall your Simple Algebra…

- If the same quantity exists in the Numerator and Denominator of a fraction, they "cancel each other out"
  - Leaving us with a number (F) that represents Group Differences!

The F-Ratio

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's} + \text{error} + \text{group differences}}{\text{ID's} + \text{error}}$$

Analysis of Variance F-Ratio



- If F=1…
- As F increases…
- How do you know if F is "big enough" to considered significant?
  - How do you know a t-test is significant??

The F-Ratio

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's} + \text{error} + \text{group differences}}{\text{ID's} + \text{error}}$$

Confidence Intervals with the F-test

- CI’s for comparing two groups are straightforward and intuitive
- CI’s for “Omnibus” differences are less so
  - Effect size calculations exist, but less intuitive interpretation..
- Stay tuned for discussions about post-hoc tests, and how they can sometimes help
- Plots will also be very informative

IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups

IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups

Sum of Squared Deviations from the Mean

IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups

Like in a t-test, each F-test has df values for significance testing

IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups

MS is the Variance Statistic for ANOVA—calculated with SS & df

IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups

The "F" statistic is another word for the F-ratio

IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups

...and p values tell us the significance level of the ratio

This is what it looks like...

	df	SS	MS	F	p
Between Groups	##	##	###	#.#	.##
Within Groups (error)	##	##	###		

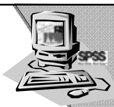
This is where it comes from (Independent Measures Designs)

$df_{within} = N - k$

This is where it comes from (Independent Measures Designs)

F-tables provide a p value for a given F-statistic, using $df_{between}$ (numerator) and df_{within} (denominator).

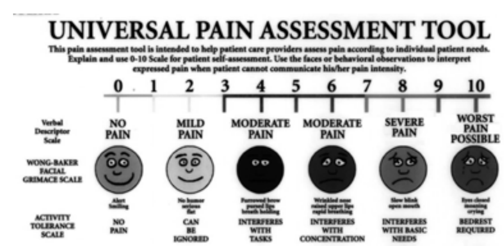
Example 1



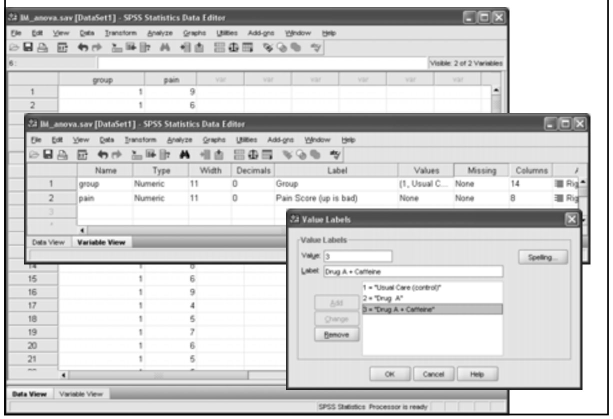
- Compare Pain Ratings of Patients in Randomized Clinical Trial
  - Usual Care (control)
  - Pain Medication
  - Pain Medication + Caffeine
- Simplest of ANOVA Models, with ONE Independent Factor (Treatment Group)

Design: One-way ANOVA

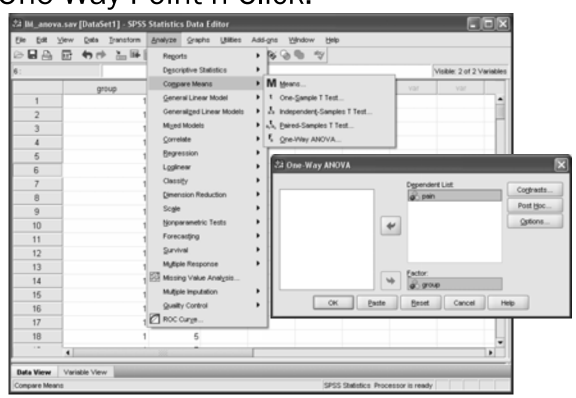
- Compare Three Groups on Pain Assessment



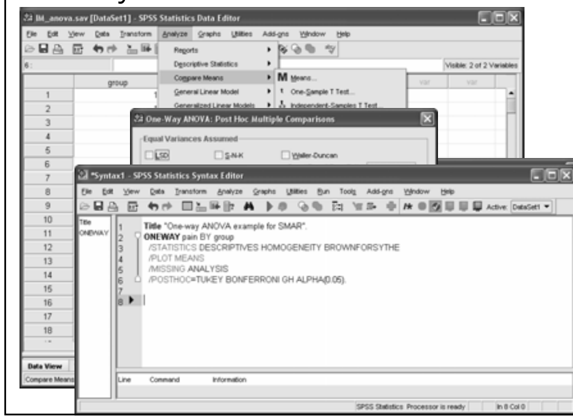
The Data:



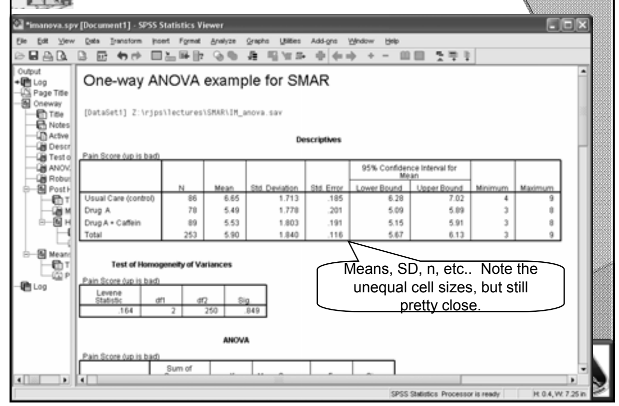
One-Way Point-n-Click:



One-Way Point-n-Click:



Results



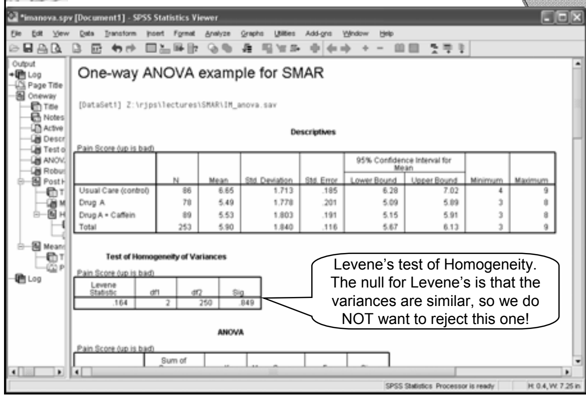
Descriptives — Pain Score (up is bad)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean: Lower Bound	95% Confidence Interval for Mean: Upper Bound	Minimum	Maximum
Usual Care (control)	86	6.65	1.713	.185	6.28	7.02	4	9
Drug A	78	5.49	1.778	.201	5.09	5.89	3	8
Drug A + Caffein	89	5.53	1.803	.191	5.15	5.91	3	8
Total	253	5.90	1.840	.116	5.67	6.13	3	9

Test of Homogeneity of Variances — Pain Score (up is bad)

Levene Statistic	df1	df2	Sig.
.164	2	250	.849

Results



Descriptives

Pain Score (up is bad)

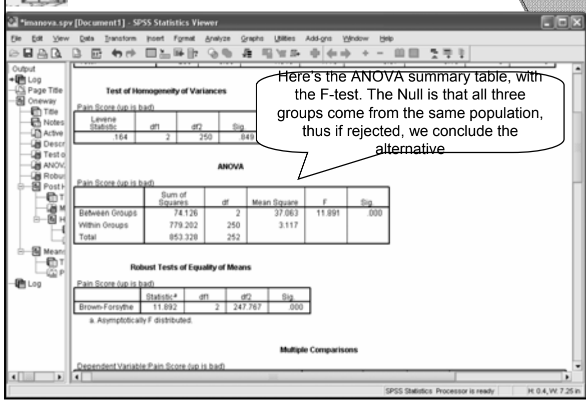
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
Usual Care (control)	86	6.65	1.713	.185	6.28	7.02	4	9
Drug A	78	5.49	1.778	.201	5.09	5.89	3	8
Drug A + Caffein	89	5.53	1.803	.191	5.15	5.91	3	8
Total	253	5.90	1.840	.116	5.67	6.13	3	9

Test of Homogeneity of Variances

Pain Score (up is bad)

Levene Statistic	df1	df2	Sig.
.164	2	250	.849

Results



ANOVA

Pain Score (up is bad)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	74.126	2	37.063	11.891	.000
Within Groups	779.202	250	3.117		
Total	853.328	252			

Robust Tests of Equality of Means

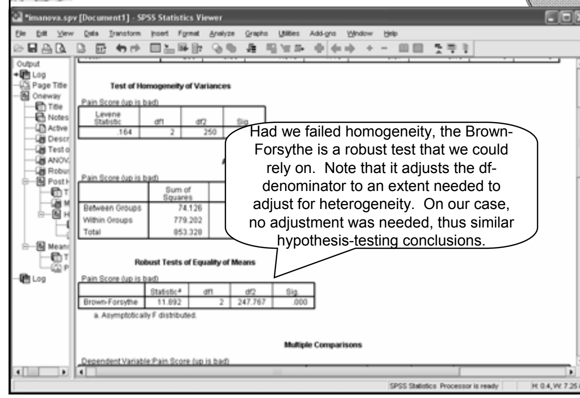
Pain Score (up is bad)

	Statistic[a]	df1	df2	Sig.
Brown-Forsythe	11.892	2	247.767	.000

a. Asymptotically F distributed.

Multiple Comparisons

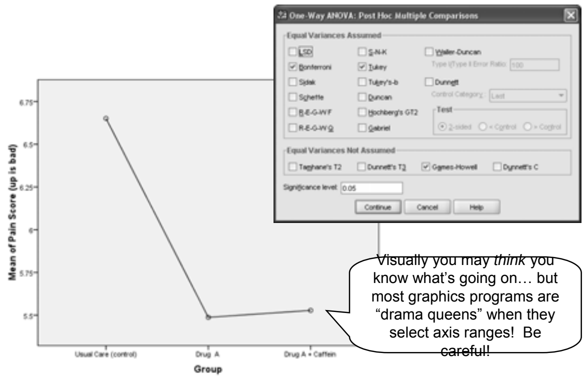
Results



Satisfied?

- We rejected the Omnibus F-test, concluding that the three groups must be different… All done?
- Usually pairwise comparisons are of interest too
  - Post-Hoc (compares all pairs)
    - Different choices available
  - A-Priori Contrasts (hypothesis-specific subset of comparisons)
    - Different choices available

Recall our earlier analytic choices…:



Post-hoc Pairwise Comparisons

Multiple Comparisons

Dependent Variable:Pain Score (up is bad)

	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Usual Care (control)	Drug A	1.164*	.276	0.000	.51	1.81
		Drug A + Caffein	1.123*	.267	0.000	.49	1.75
	Drug A	Usual Care (control)	-1.164*	.276	.000	-1.81	-.51
		Drug A + Caffein	-.041	.274	0.988	-.69	.60
	Drug A + Caffein	Usual Care (control)	-1.123*	.267	.000	-1.75	-.49
		Drug A	.041	.274	.988	-.60	.69
Bonferroni	Usual Care (control)	Drug A	1.164*	.276	.000	.50	1.83
		Drug A + Caffein	1.123*	.267	.000	.48	1.77
	Drug A	Usual Care (control)	-1.164*	.276	.000	-1.83	-.50
		Drug A + Caffein	-.041	.274	1.000	-.70	.62
	Drug A + Caffein	Usual Care (control)	-1.123*	.267	.000	-1.77	-.48
		Drug A	.041	.274	1.000	-.62	.70
Games-Howell	Usual Care (control)	Drug A	1.164*	.273	.000	.52	1.81
		Drug A + Caffein	1.123*	.266	.000	.49	1.75
	Drug A	Usual Care (control)	-1.164*	.273	.000	-1.81	-.52
		Drug A + Caffein	-.041	.278	.988	-.70	.62
	Drug A + Caffein	Usual Care (control)	-1.123*	.266	.000	-1.75	-.49
		Drug A	.041	.278	.988	-.62	.70

*. The mean difference is significant at the 0.05 level.

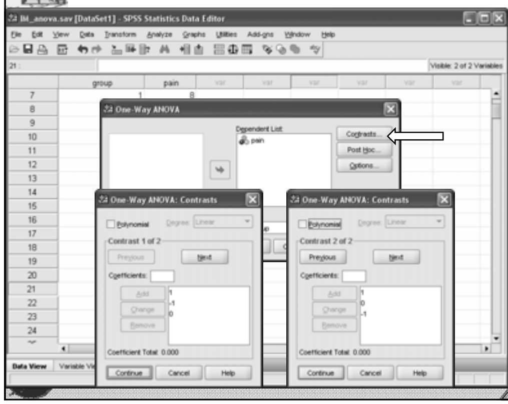
If we had a-priori contrasts?

- We chose all possible pairs to compare post-hoc, adjusting for the number of comparisons
  - UC vs. Treatment A
  - UC vs. Treatment A + Caffeine
  - Treatment A vs. Treatment A+Caffeine
- If we had a more specific set of comparisons that we wanted to make *A-priori*, we could have more statistical power, at the expense of unnecessary comparisons.

In this example?

- May make sense to compare Usual care to either of the novel Treatments, but not to compare the two novel treatments?
- "Simple" contrasts, with a reference category (usual care)
  - Usual Care vs. Treatment A
  - Usual Care vs. Treatment A + Caffeine

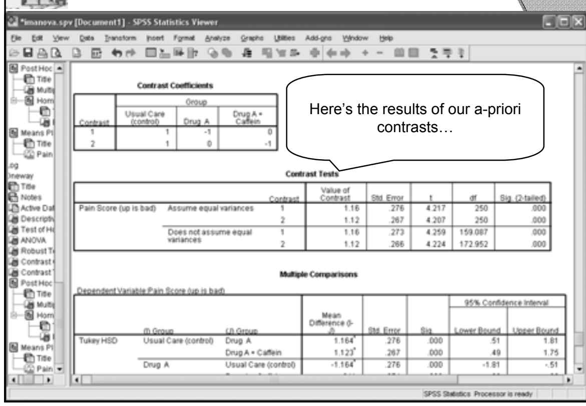
Back to our data…



Cake anyone?

- You can't have your cake, and eat it too!
- Good Science dictates that you either HAVE a-priori contrasts, or you DON'T!
  - Contrasts are theory-driven, not something that you do "after the fact"
  - Post-hoc tests are more appropriate if you want all possible pairs of comparisons

Results of the Contrasts



Contrast Coefficients

Contrast	Group: Usual Care (control)	Drug A	Drug A + Caffein
1	1	-1	0
2	1	0	-1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Pain Score (up is bad)	Assume equal variances	1	1.16	.276	4.217	250	.000
		2	1.12	.267	4.207	250	.000
	Does not assume equal variances	1	1.16	.273	4.259	159.087	.000
		2	1.12	.266	4.224	172.952	.000

Multiple Comparisons

Dependent Variable:Pain Score (up is bad)

	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Tukey HSD	Usual Care (control)	Drug A	1.164*	.276	.000	.51	1.81
		Drug A + Caffein	1.123*	.267	.000	.49	1.75
	Drug A	Usual Care (control)	-1.164*	.276	.000	-1.81	-.51

Next Time: Two-Factor ANOVAS

- What if you want to compare 2+ groups on MORE THAN one factor?
  - Effect of subjects' gender *and* Treatment on BMD?
  - Effect of Novel Treatment (vs. control) *and* Implementation Schedule (two types) on Countermeasure's Effectiveness?
  - Effect of Suit Pressure (3 settings) *and* Glove Design (2 types) on EVA performance?

SMAR Session 3

ANOVA

(CONTINUED)

Recap

- Analysis of Variance (ANOVA) examines variability between groups, relative to within groups, to determine whether there's evidence that the groups are not from the same population
- Analysis focuses on *variance*, but interpretation is about *mean's*
- One-way ANOVA compares more than two groups.
  - Similar to a t-test, but for 3, 4, 5+ groups

Recap

- ANOVA assumes
  - Random samples from the population
  - Sufficiently large enough n to detect effects, distributed evenly among groups
  - Similar variability among groups (Homogeneity of Variance)
- We should examine our data and test our assumptions
  - Sometimes we need to consider data transformations to meet these assumptions
  - Sometimes we need to rely on robust alternatives to the typical ANOVA statistic

Recap

- ANOVA results summarized in a ANOVA table, with an "Omnibus F-statistic" and p-value
  - Represents the ratio of between/within variability
  - If significant, reject the null hypothesis that the groups are from the same population
- Researchers typically follow-up a significant F-ratio with either
  - Post Hoc tests
  - A-Priori Contrasts

Today… Multifactorial ANOVA

- What if you want to compare 2+ groups on MORE THAN one factor?
  - Effect of subjects' gender *and* Treatment on BMD?
  - Effect of Novel Treatment (vs. control) *and* Implementation Schedule (two types) on Countermeasure's Effectiveness?
  - Effect of Suit Pressure (3 settings) *and* Glove Design (2 types) on EVA performance?
- Still working with completely Independent Measures Designs
  - Subjects in one "cell" are not also in any other "cell" of the design

Table Representation of Experimental Two-Factor Designs

3 x 2 design Study n=120	No Drug	Low Dose	High Dose
Drug A	n=20	n=20	n=20
Drug B	n=20	n=20	n=20

3 x 3 design Study n=180	No Drug	Low Dose	High Dose
No Therapy	n=20	n=20	n=20
Therapy A	n=20	n=20	n=20
Therapy B	n=20	n=20	n=20

2 x 2 design Study n=78	Control	Intervention
Males	n=20	n=18
Females	n=19	n=21

More Complicated Designs:

- ANOVA can handle 3, 4, 5, or even more factors!
  - "k" is the common notation for number of factors in an ANOVA design
- But be careful what you ask for… stay tuned!

2x2x2 design Study n=152	Placebo		Experimental Drug	
	Chronic Use	Acute Administration	Chronic Use	Acute Administration
Males	n=20	n=18	n=19	n=17
Females	n=19	n=21	n=18	n=20

Main Effects and Interactions

- Main Effects
  - One per factor…an F-statistic evaluating the impact of each factor in the model
    - Gender effect on performance (M/F diffs?)
    - Race/ethnicity effect on performance
- Interaction Effects
  - One per interaction… an F-statistic evaluating how two (or more) factors interact with one another to affect the outcome
    - Gender "by" Race/Ethnicity interactive effects on performance
    - More complex…often more interesting!

Interactions…

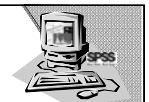
- Interaction effects are often the most interesting, but can be tricky to understand at first
- We describe them in terms of how many factors are involved
  - “Two-way” means two factors work together to explain the observed difference
  - “Three-way” means that three factors tell the story
  - “Four-way” means that you’d better have some pain relievers nearby!

The Usual Assumptions…

- Random Sample
- Roughly Equal n Per Cell
- Continuously Scaled outcome (Performance Gains) follows Normal Distribution
- Homogeneity of Variance

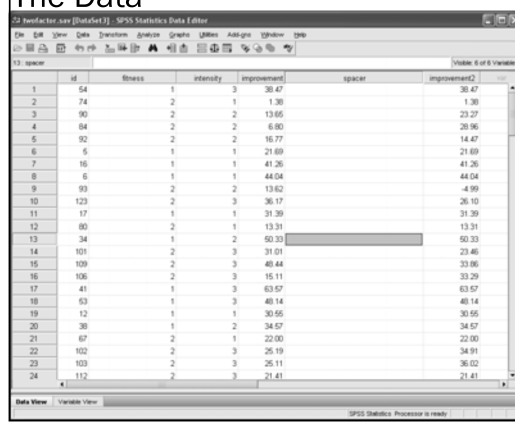
2 x 3 design Study n=116	Exercise Intensity		
	Low	Med	High
Couch Potatoes	n=20	n=19	n=18
Fit Individuals	n=18	n=19	n=22

Two-Factor Example



- Compare Performance Gains Following Exercise Intervention by Subjects’ Initial Fitness Status
  - Subjects’ Current Fitness Level Upon Enrollment in Study
    - Couch Potatoes, Fit Individuals
  - Intensity of Exercise Program
    - Low, Medium, High
- 2 x 3 ANOVA

The Data



twofactor.sav [DataSet3] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

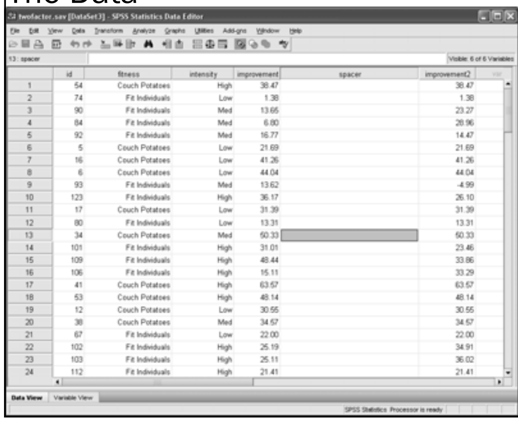
13 : spacer — Visible: 6 of 6 Variables

	id	fitness	intensity	improvement	spacer	improvement2
1	54	1	3	38.47		38.47
2	74	2	1	1.38		1.38
3	90	2	2	13.65		23.27
4	84	2	2	6.80		28.96
5	92	2	2	16.77		14.47
6	5	1	1	21.69		21.69
7	16	1	1	41.26		41.26
8	6	1	1	44.04		44.04
9	93	2	2	13.62		-4.99
10	123	2	3	36.17		26.10
11	17	1	1	31.39		31.39
12	80	2	1	13.31		13.31
13	34	1	2	50.33		50.33
14	101	2	3	31.01		23.46
15	109	2	3	48.44		33.86
16	106	2	3	15.11		33.29
17	41	1	3	63.57		63.57
18	63	1	3	48.14		48.14
19	12	1	1	30.55		30.55
20	38	1	2	34.57		34.57
21	67	2	1	22.00		22.00
22	102	2	3	25.19		34.91
23	103	2	3	25.11		36.02
24	112	2	3	21.41		21.41

Data View Variable View

SPSS Statistics Processor is ready

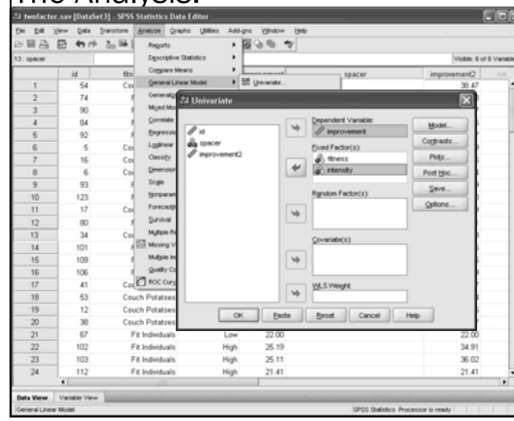
The Data



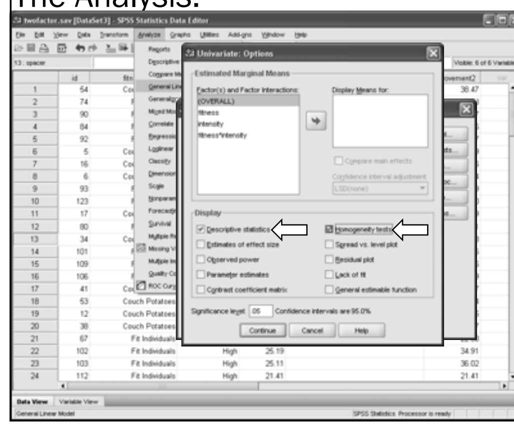
First Example

- Use “Improvement” as our outcome variable

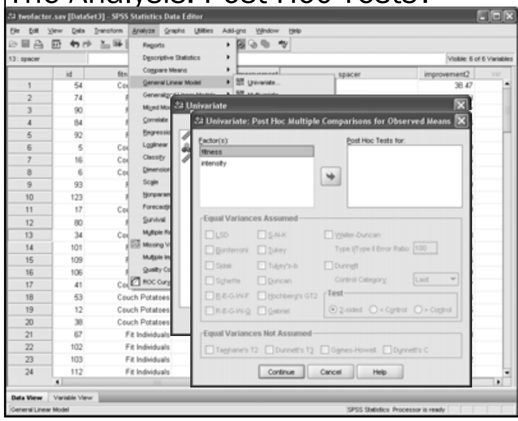
The Analysis:



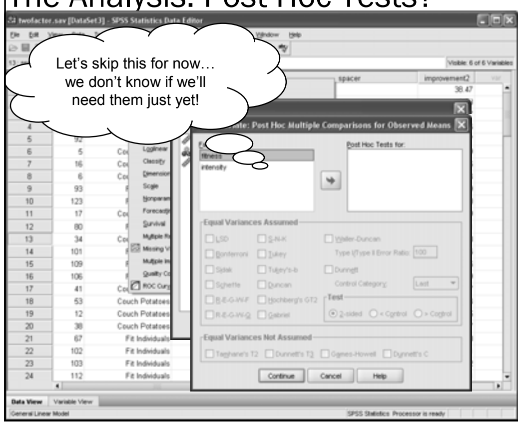
The Analysis:



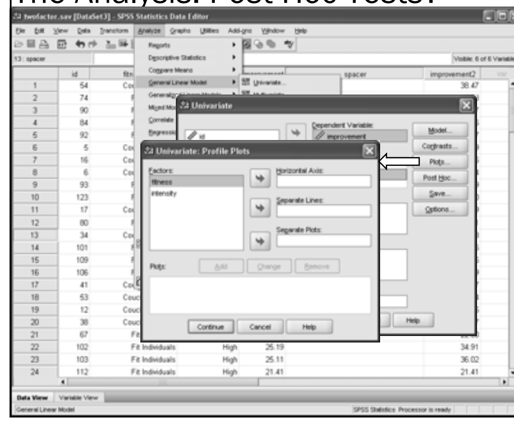
The Analysis: Post-Hoc Tests?



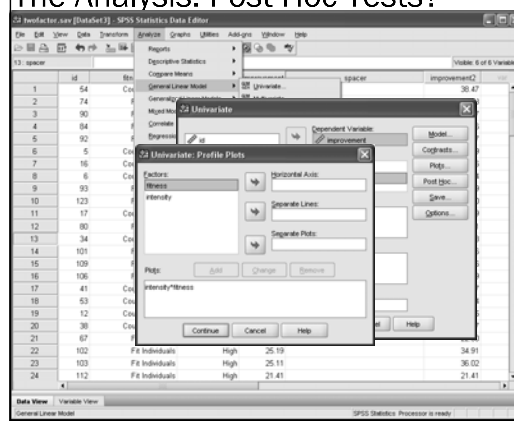
The Analysis: Post-Hoc Tests?



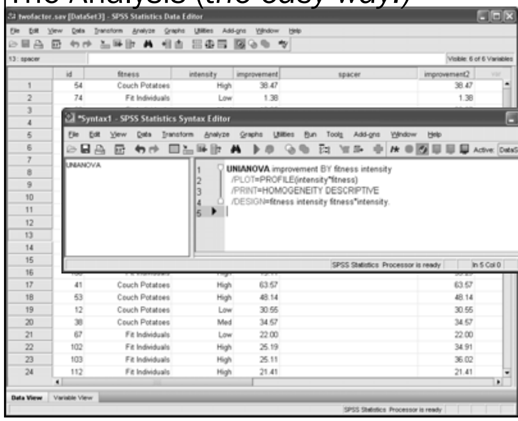
The Analysis: Post-Hoc Tests?



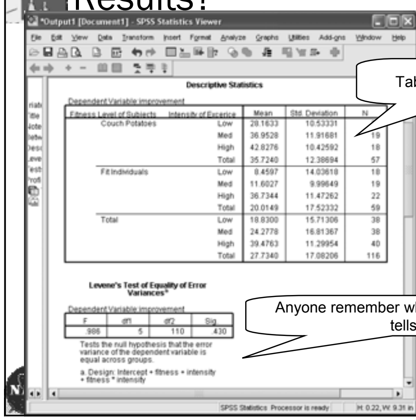
The Analysis: Post-Hoc Tests?



The Analysis (*the easy way!*)



Results?



Descriptive Statistics

Dependent Variable: improvement

Fitness Level of Subjects	Intensity of Excerice	Mean	Std. Deviation	N
Couch Potatoes	Low	28.1633	10.53331	19
	Med	36.9528	11.91681	19
	High	42.8276	10.42592	18
	Total	35.7240	12.38694	57
Fit Individuals	Low	8.4597	14.03618	18
	Med	11.6027	9.99649	19
	High	36.7344	11.47262	22
	Total	20.0149	17.52332	59
Total	Low	18.8300	15.71306	38
	Med	24.2778	16.81367	38
	High	39.4763	11.29954	40
	Total	27.7340	17.08206	116

Levene's Test of Equality of Error Variances[a]

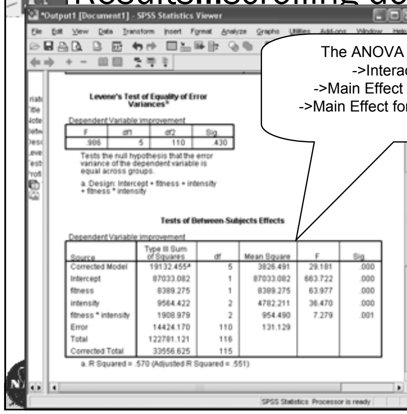
Dependent Variable: improvement

F	df1	df2	Sig.
.986	5	110	.430

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + fitness + intensity + fitness * intensity

Results…scrolling down.



Levene's Test of Equality of Error Variances[a]

Dependent Variable: improvement

F	df1	df2	Sig.
.986	5	110	.430

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + fitness + intensity + fitness * intensity

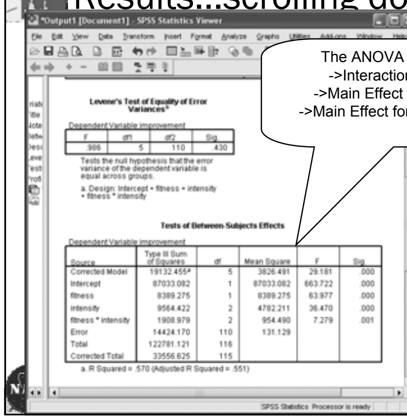
Tests of Between-Subjects Effects

Dependent Variable: improvement

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	19132.455[a]	5	3826.491	29.181	.000
Intercept	87033.082	1	87033.082	663.722	.000
fitness	8389.275	1	8389.275	63.977	.000
intensity	9564.422	2	4782.211	36.470	.000
fitness * intensity	1908.979	2	954.490	7.279	.001
Error	14424.170	110	131.129		
Total	122781.121	116			
Corrected Total	33556.625	115			

a. R Squared = .570 (Adjusted R Squared = .551)

Results…scrolling down.



Levene's Test of Equality of Error Variances[a]

Dependent Variable: improvement

F	df1	df2	Sig.
.986	5	110	.430

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + fitness + intensity + fitness * intensity

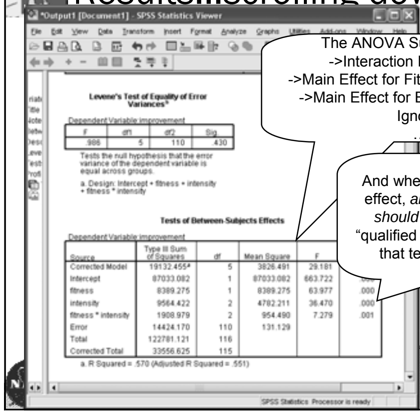
Tests of Between-Subjects Effects

Dependent Variable: improvement

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	19132.455[a]	5	3826.491	29.181	.000
Intercept	87033.082	1	87033.082	663.722	.000
fitness	8389.275	1	8389.275	63.977	.000
intensity	9564.422	2	4782.211	36.470	.000
fitness * intensity	1908.979	2	954.490	7.279	.001
Error	14424.170	110	131.129		
Total	122781.121	116			
Corrected Total	33556.625	115			

a. R Squared = .570 (Adjusted R Squared = .551)

Results…scrolling down.


*Output1 [Document1] - SPSS Statistics Viewer

Levene's Test of Equality of Error Variances[a]

Dependent Variable:improvement

F	df1	df2	Sig.
.986	5	110	.430

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + fitness + intensity + fitness * intensity

Tests of Between-Subjects Effects

Dependent Variable:improvement

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	19132.455[a]	5	3826.491	29.181	[illegible]
Intercept	87033.082	1	87033.082	663.722	[illegible]
fitness	8389.275	1	8389.275	63.977	.000
intensity	9564.422	2	4782.211	36.470	.000
fitness * intensity	1908.979	2	954.490	7.279	.001
Error	14424.170	110	131.129		
Total	122781.121	116			
Corrected Total	33556.625	115			

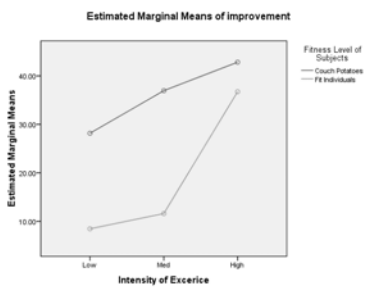
a. R Squared = .570 (Adjusted R Squared = .551)

SPSS Statistics Processor is ready

The ANOVA Summary Table:
->Interaction Effect? YES!!!
->Main Effect for Fitness Level? Ignore!
->Main Effect for Exercise Intensity? Ignore!
…

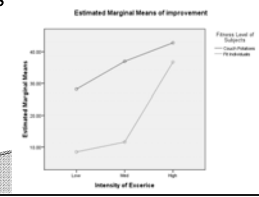
And when there's an Interaction effect, *all Main Effects involved should be ignored.* They are "qualified by" an Interaction Effect that tells more of the story.

Let's look at the plot for kicks



Interaction Plots

- Lines that aren't parallel suggest an Interaction Effect
  - One line is different from the other
- Sometimes the lines cross over one another, forming an "X," "zig-zag" or other
  - Dramatically different effects



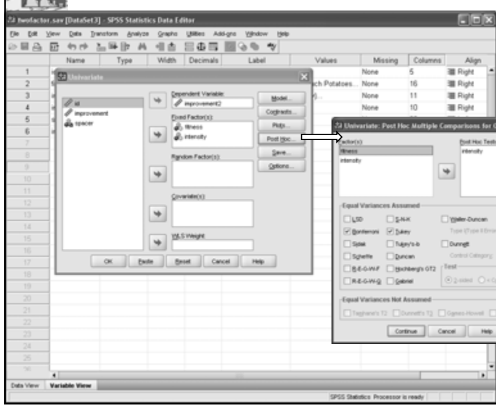
We have an interaction, Now What?

- A-priori hypotheses rule the day!
- If you had 'em, test 'em!
- If not (more typically), then many will argue that you are finished with analyzing the data
- Plot the means (albeit better than SPSS does by default), report the Interaction effect, and explain.
- Others argue that you are justified for running additional analyses to look for potential differences in one factor, nested within levels of the other

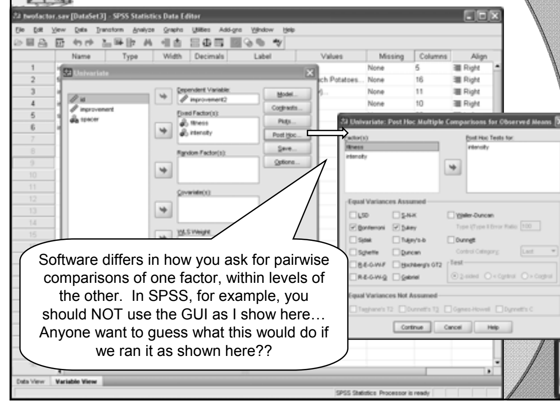
Our Study?

- Let's assume we thought there would be a difference in effects of low, medium, high intensity by fitness levels
- Let's further assume that we did not have a-priori hypotheses begging for specific contrasts, but rather wanted to follow-up with whatever post-hoc tests we are justified at running

How *not* to do this with SPSS?

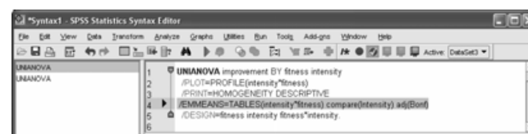


How *not* to do this with SPSS?

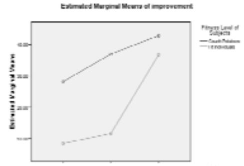


Simple Effects

- To get pairwise comparisons of Intensity(Low, Medium, High) within levels of Fitness (Couch Potato, Fit Individuals), we need to use Syntax (code) as shown below



Results…



Pairwise Comparisons

Dependent Variable: improvement

Fitness Level of Subjects	(I) Intensity of Exercise	(J) Intensity of Exercise	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a] Lower Bound	Upper Bound
Couch Potatoes	Low	Med	-8.789	3.669	.055	-17.708	.129
		High	-14.664*	3.720	.000	-23.709	-5.620
	Med	Low	8.789	3.669	.055	-.129	17.708
		High	-5.875	3.766	.365	-15.032	3.282
	High	Low	14.664*	3.720	.000	5.620	23.709
		Med	5.875	3.766	.365	-3.282	15.032
Fit Individuals	Low	Med	-3.143	3.766	1.000	-12.300	6.014
		High	28.270*	3.639	.000	37.123	19.427
	Med	Low	3.143	3.766	1.000	-6.014	12.300
		High	-25.132*	3.586	.000	-33.851	-16.413
	High	Low	28.270*	3.639	.000	19.427	37.123
		Med	25.132*	3.586	.000	16.413	33.851

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

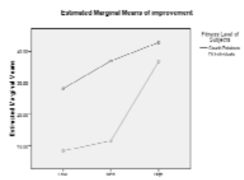
*. The mean difference is significant at the .050 level.

Write-up?

…Our analysis revealed a significant Fitness level by Exercise Intensity Interaction Effect (p< 0.01). Mean data reveal how the differential effect of Low, Medium, or High Intensity Exercise programs affected Couch Potatoes differently than Fit Individuals (see fig).

Given the significant interaction, we also ran simple effects contrasts (Bonferroni adjusted) comparing the pairwise effects of Low, Medium and High Intensity exercise interventions within Fitness level. Couch Potatoes randomized to the High-Intensity intervention showed significantly greater gains relative to Couch Potatoes in the Low-Intensity condition (p<.01). Neither of these groups differed from the Medium-Intensity Couch Potatoes group, though the Low-Intensity Couch Potatoes were marginally significantly different from their Medium-Intensity counterparts (p=.055).

Write-up?



…Our analysis revealed a significant Fitness level by Exercise Intensity Interaction Effect (p< 0.01). Mean data reveal how the differential effect of Low, Medium, or High Intensity Exercise programs affected Couch Potatoes differently than Fit Individuals (see fig).

(You could stop here and discuss some theory explaining why this may be the case. You could also continue with the simple effects if you think that it further illuminates important findings.)

Write-up?

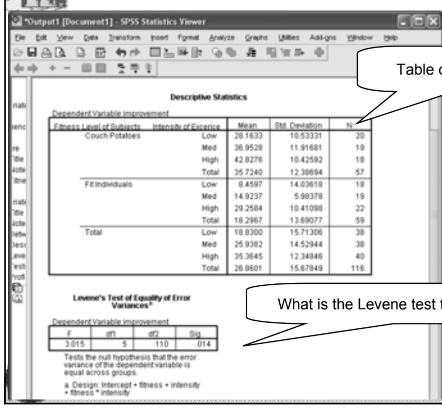
… Given the significant interaction, we also ran simple effects contrasts comparing the pairwise effects of Low, Medium and High Intensity exercise interventions within Fitness level. Couch Potatoes randomized to the High-Intensity intervention showed significantly greater gains relative to Couch Potatoes in the Low-Intensity condition (p<.01). Neither of these groups differed from the Medium-Intensity Couch Potatoes group, though the Low-Intensity Couch Potatoes were marginally significantly different from their Medium-Intensity counterparts (p=.055).

In contrast, the pairwise comparisons with the Fit Individuals groups revealed significant and clear benefits of those randomized to High-Intensity exercise relative to Low (p<.01), and those in the High versus Medium Intensity groups (p<.01). The marginal effect observed among Couch Potato Low versus High subjects was not evident in Fit Individuals randomized to either of these exercise interventions.

Next?

- Let's try it again using "Improvement2" instead of "Improvement" as our observed results.
- This is for illustration purposes—pretend like these were our data instead of the earlier results…
  - The analysis set-up is the same.

Results?



Descriptive Statistics

Dependent Variable: improvement

Fitness Level of Subjects	Intensity of Excerice	Mean	Std. Deviation	N
Couch Potatoes	Low	28.1633	10.53331	20
	Med	36.9528	11.91681	19
	High	42.8276	10.42592	18
	Total	35.7240	12.38694	57
Fit Individuals	Low	8.4597	14.03618	18
	Med	14.9237	5.98378	19
	High	29.2584	10.41098	22
	Total	18.2967	13.69077	59
Total	Low	18.8300	15.71306	38
	Med	25.9382	14.52944	38
	High	35.3645	12.34846	40
	Total	26.8601	15.67849	116

Levene's Test of Equality of Error Variances[a]

Dependent Variable: improvement

F	df1	df2	Sig.
3.015	5	110	.014

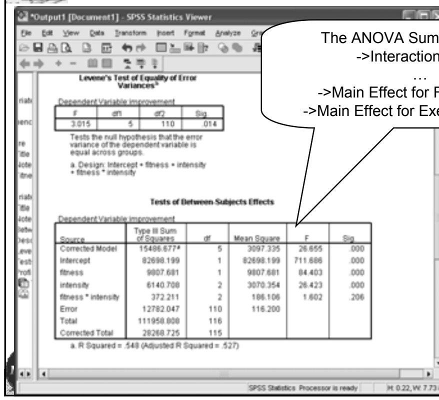
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + fitness + intensity + fitness * intensity

Table of means & sd…

What is the Levene test telling us this time?

Results…scrolling down



Levene's Test of Equality of Error Variances[a]

Dependent Variable: improvement

F	df1	df2	Sig.
3.015	5	110	.014

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + fitness + intensity + fitness * intensity

Tests of Between Subjects Effects

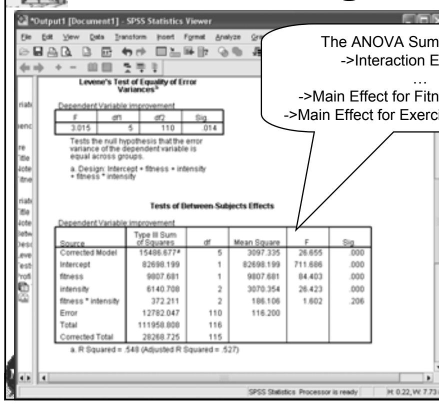
Dependent Variable: improvement

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15486.677[a]	5	3097.335	26.655	.000
Intercept	82698.199	1	82698.199	711.686	.000
fitness	9807.681	1	9807.681	84.403	.000
intensity	6140.708	2	3070.354	26.423	.000
fitness * intensity	372.211	2	186.106	1.602	.206
Error	12782.047	110	116.200		
Total	111958.808	116			
Corrected Total	28268.725	115			

a. R Squared = .548 (Adjusted R Squared = .527)

The ANOVA Summary Table:
->Interaction Effect
…
->Main Effect for Fitness Level
->Main Effect for Exercise Intensity

Results…scrolling down



Levene's Test of Equality of Error Variances[a]

Dependent Variable: improvement

F	df1	df2	Sig.
3.015	5	110	.014

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + fitness + intensity + fitness * intensity

Tests of Between Subjects Effects

Dependent Variable: improvement

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15486.677[a]	5	3097.335	26.655	.000
Intercept	82698.199	1	82698.199	711.686	.000
fitness	9807.681	1	9807.681	84.403	.000
intensity	6140.708	2	3070.354	26.423	.000
fitness * intensity	372.211	2	186.106	1.602	.206
Error	12782.047	110	116.200		
Total	111958.808	116			
Corrected Total	28268.725	115			

a. R Squared = .548 (Adjusted R Squared = .527)

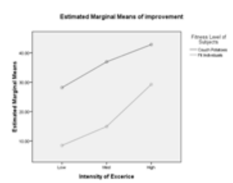
The ANOVA Summary Table:
->Interaction Effect NO
…
->Main Effect for Fitness Level YES!
->Main Effect for Exercise Intensity YES!

Interpreting Multi-Factorial ANOVA Results

- If you have an Interaction Effect, Start There!
  - All main effects involved in a significant interaction are qualified effects anyway—they don't tell the whole story
  - This was the case in our earlier example
- If you do not have Interaction Effects, Interpret whatever Main Effects you have
  - This is the case in our current example
    - Post-hoc or Contrast effects may be useful now?

Back to our Example…

- Main effect for Fitness Level
  - Couch Potatoes improved more than Fit Individuals
- Main effect for Intensity of the Exercise
  - Looks like increasing intensity produced greater benefits overall
  - Follow-up?



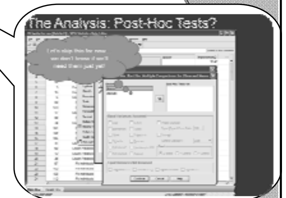
Tests of Between-Subjects Effects

Dependent Variable:improvement

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15486.677[a]	5	3097.335	26.655	.000
Intercept	82698.199	1	82698.199	711.686	.000
fitness	9807.681	1	9807.681	84.403	.000
intensity	6140.708	2	3070.354	26.423	.000
fitness * intensity	372.211	2	186.106	1.602	.206
Error	12782.047	110	116.200		
Total	111958.808	116			
Corrected Total	28268.725	115			

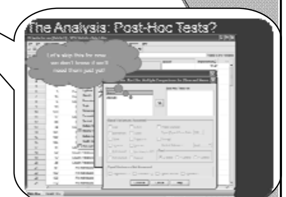
a. R Squared = .548 (Adjusted R Squared = .527)

Remember when we skipped post-hoc tests?



- Now that we know we don't have an Interaction effect, it's time to consider post-hoc tests, if you desire
  - Compare All Intensity Levels Pairwise, collapsing across Fitness Level
- Why not contrast comparisons of the levels of one factor "within" the other?

Remember when we skipped post-hoc tests?



- Now that we know we don't have an Interaction effect, it's time to consider post-hoc tests, if you desire
  - Compare All Intensity Levels Pairwise, collapsing across Fitness Level
- Why not contrast comparisons of the levels of one factor "within" the other?
  - No Evidence that the differences between Low, Med, High are any different in Couch Potatoes versus Fit People!

Ok to re-run this as a Oneway?

- We know that there's no interaction effect, so can we just run it as a oneway and get the pairwise comparisons that way?
  - Nope. That would ignore the variance structure in the data, *and* inflates Type I error
  - This was a multifactorial design from the start, so stick with your multifactorial analysis

Back to our analysis:

- Generate posthoc tests for the *Intensity* factor (GUI or Syntax, your preference)
- And remember that we're now, essentially, averaging across levels of Subject Fitness (the Couch Potatoes and Fit Individuals), but we are doing so in a multivariate context


```
UNIANOVA improvement BY fitness intensity
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /POSTHOC=intensity(TUKEY BONFERRONI)
  /PLOT=PROFILE(intensity*fitness)
  /PRINT=HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=fitness intensity fitness*intensity
```

Pairwise Comparisons

Multiple Comparisons

Dependent Variable improvement

	(I) Intensity of Excerice	(J) Intensity of Excerice	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Low	Med	-7.1082*	2.47302	.013	-12.9838	-1.2327
		High	-16.5345*	2.44191	.000	-22.3362	-10.7329
	Med	Low	7.1082*	2.47302	.013	1.2327	12.9838
		High	-9.4263*	2.44191	.001	-15.2279	-3.6247
	High	Low	16.5345*	2.44191	.000	10.7329	22.3362
		Med	9.4263*	2.44191	.001	3.6247	15.2279
Bonferroni	Low	Med	-7.1082*	2.47302	.015	-13.1205	-1.0960
		High	-16.5345*	2.44191	.000	-22.4712	-10.5979
	Med	Low	7.1082*	2.47302	.015	1.0960	13.1205
		High	-9.4263*	2.44191	.001	-15.3629	-3.4897
	High	Low	16.5345*	2.44191	.000	10.5979	22.4712
		Med	9.4263*	2.44191	.001	3.4897	15.3629

Based on observed means.
The error term is Mean Square(Error) = 116.200.
*. The mean difference is significant at the .05 level.

BTW... if you ran as a Oneway?

- Similar conclusions, but note that Low vs. Med is no longer significant??

Multiple Comparisons

Dependent Variable improvement

	(I) Intensity of Excerice	(J) Intensity of Excerice	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Low	Med	-7.10824	3.26540	.080	-14.8635	.6471
		High	-16.53454*	3.22432	.000	-24.1923	-8.8768
	Med	Low	7.10824	3.26540	.080	-.6471	14.8635
		High	-9.42630*	3.22432	.012	-17.0840	-1.7686
	High	Low	16.53454*	3.22432	.000	8.8768	24.1923
		Med	9.42630*	3.22432	.012	1.7686	17.0840
Games-Howell	Low	Med	-7.10824	3.47171	.108	-15.4128	1.1963
		High	-16.53454*	3.21084	.000	-24.2226	-8.8465
	Med	Low	7.10824	3.47171	.108	-1.1963	15.4128
		High	-9.42630*	3.06064	.008	-16.7493	-2.1033
	High	Low	16.53454*	3.21084	.000	8.8465	24.2226
		Med	9.42630*	3.06064	.008	2.1033	16.7493

*. The mean difference is significant at the 0.05 level.

Next time

- Quick Discussion of why, in general, we advise against 3-factor, 4-factor…k-factorial models
- Move into Repeated Measures designs
  - Pre, Post1, Post2, Post3

SMAR Session 4

ANOVA

(CONTINUED)

Recap—Independent Measures ANOVA

- IM Analysis of Variance (ANOVA) examines variability between groups, relative to within groups, to determine whether there's evidence that the groups are not from the same population
- Analysis focuses on *variance*, but interpretation is about *mean's*
- One-way ANOVA compares more than two groups, defined by a single Factor
- Multi-Factorial considers additional factors

Recap—Independent Measures ANOVA

- With Multiple Factors:
  - Main Effects
  - Interaction Effects
- IM ANOVA assumes
  - Random samples from the population
  - Sufficiently large enough n to detect effects, distributed evenly among groups
  - Similar distributions of variability among groups (Homogeneity of Variance)

Recap—Independent Measures ANOVA

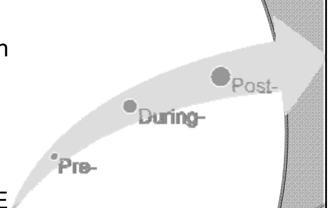
- General Strategy is to Interpret Significant Interactions if you have them
  - Main Effects only tell part of the story
  - Simple Effects can help further
- If no Interactions, Interpret Main Effects
  - Post-Hoc or Contrasts Available for Pairwise Comparisons

Today: Repeated Measures ANOVA

- With RM ANOVA, we consider measuring the SAME subjects at different times, or under different conditions, to see if something changes over time, or between the different conditions
  - Ex. Does performance decrease in response to time spent in microgravity?
  - Ex. Does bone mass decrease as subjects age?
  - Ex. Compare Subjects' Ratings of 'XYZ' when taking Placebo, versus Drug A, versus Drug B, with adequate washout periods between drugs
  - Ex. Compare PRE to Post1 and Post2…

Repeated Measures Designs

- Only one sample
- Differences based on time, or condition
- Using the SAME subjects time after time
- Measuring the SAME outcome each time
- Looking for changes…



Repeated Measures ANOVA

- Same people… no "individual differences" in the F-ratio
- More powerful statistics

- The F-Ratio represents:

$$F = \frac{\text{Variability among times (or conditions)}}{\text{Variability within the sample}} = \frac{\text{error + time (or condition) differences}}{\text{error}}$$

Repeated Measures ANOVA

- Same people… no "individual differences" in the F-ratio
- More powerful statistics

$$F = \frac{\text{Variability among times (or conditions)}}{\text{Variability within the sample}} = \frac{\cancel{\text{error}} \text{ + time (or condition) differences}}{\cancel{\text{error}}}$$

Assumptions for RM ANOVA

- Same as for IM-ANOVA RE Ordinal or Continuously Scaled Outcomes following the normal distribution
- Random sampling from the population with sufficient n
  - Except now only one sample…
- Homogeneity of Variance Does Apply in *purely* RM models. (only 1 group!)
- Instead, Assumption of Sphericity
  - Assume that the covariance among pairs of repeated observations are equal.

RM-ANOVA Summary Tables

- Same Concept as IM Table, but now
  - Instead of "Between Groups" effects, we have "Between Treatments" effects
  - And also "Within Treatments"
    - Consist of subject differences (among subjects)
    - And error
- One Group measured several times, thus we partition "within group" variability into that which is due to individual differences, and error.

This is where it comes from (Repeated Measures Designs)

$SS_{total} = \text{Same as IM Anova}$

$SS_{between} = \text{Same as IM Anova}$

$SS_{within} = \text{Same as IM Anova}$

$$SS_{b/t\ subjects} = \sum \frac{(\text{each person's total across treatments})^2}{k} - \frac{\left(\sum X\right)^2}{N}$$

$SS_{error} = SS_{within} - SS_{b/t\ subjects}$

k=number of times measured

$df_{total} = n-1$

$df_{between} = k-1$

$df_{within} = n-k$

$df_{b/t\ subjects} = n-1$

$df_{error} = (n-k)-(n-1)$

This is where it comes from (Repeated Measures Designs)

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

$$F = \frac{MS_{between}}{MS_{error}}$$

F-tables provide a p value for a given F-statistic, using $df_{between}$ (numerator) and df_{error} (denominator).

Example

- Compare Performance Pre, During, and Post Bed-Rest
  - Pre (time zero)
  - Three Weeks Into Bedrest
  - Six Weeks Into Bedrest (end of bedrest)
  - Three Weeks FOLLOWING Bedrest (week 9)
- Same Subjects measured 4 times
- Equal Interval between time periods

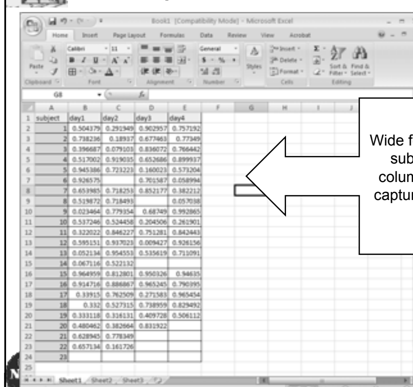
In This Example…

- Subjects (n=34) were measured on some validated performance scale four times, with equal intervals between time periods
  - Pre, Week 3, Week 6, Week 9.
  - Bedrest STOPPED at the end of Week 6
  - (Dataset created for instructional purposes)
- Prior Research has shown that these data tend to follow the normal distribution

How to Organize RM Data

- Wide Versus Long Format
  - Wide = one row per subject, with multiple column containing the multiple repeated observations
  - Long = as many rows per subject as needed, where each row contains an observation
- The Choice of Format Depends on What Software You Will Be Using
  - SPSS needs Wide
  - Stata needs Long
  - Both can convert, so for data management, use what you are comfortable with

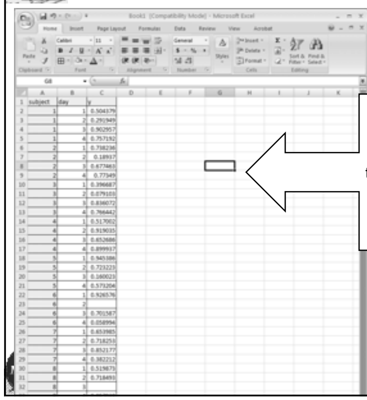
Examples of Wide Dataset



subject	day1	day2	day3	day4
1	0.504379	0.291949	0.902957	0.757192
2	0.738236	0.18937	0.677463	0.77349
3	0.396687	0.079103	0.836072	0.766442
4	0.517002	0.919035	0.652686	0.899937
5	0.945386	0.723223	0.160023	0.573204
6	0.926575		0.701587	0.058994
7	0.653985	0.718253	0.852177	0.382212
8	0.519872	0.718493		0.057038
9	0.023464	0.779354	0.68749	0.992865
10	0.537246	0.524458	0.204506	0.261901
11	0.322022	0.846227	0.751281	0.842443
12	0.595151	0.937023	0.009427	0.926156
13	0.052134	0.954553	0.535619	0.711091
14	0.067116	0.522132		
15	0.964959	0.812801	0.950326	0.94635
16	0.914716	0.886867	0.965245	0.790395
17	0.33915	0.762509	0.271583	0.965454
18	0.332	0.527315	0.738959	0.829492
19	0.333118	0.316131	0.409728	0.506112
20	0.480462	0.382664	0.831922	
21	0.628945	0.778349		
22	0.657134	0.161726		
23				

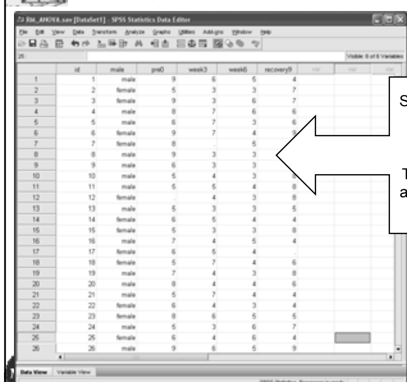
Wide format, with 1 row per subject, and as many columns as necessary to capture all of the repeated observations

Examples of Long Dataset



Long format, with as many rows per subject as needed to capture all of the repeated observations
(same data as prior slide)

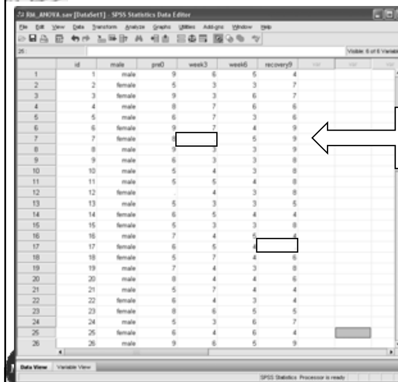
Using SPSS…



SPSS requires Wide format for Repeated Measures Designs.

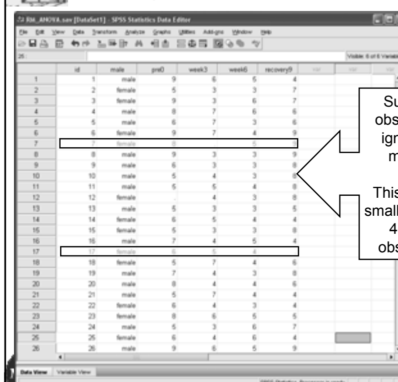
This gets tricky when there are *two* repeated measures factors…stay tuned!

Using SPSS…



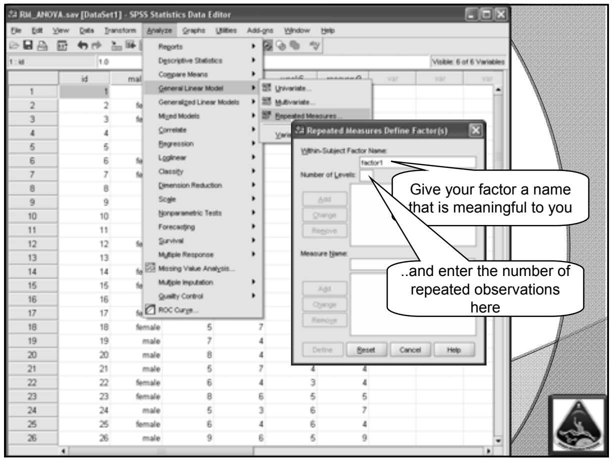
Note the occasional missing observation…

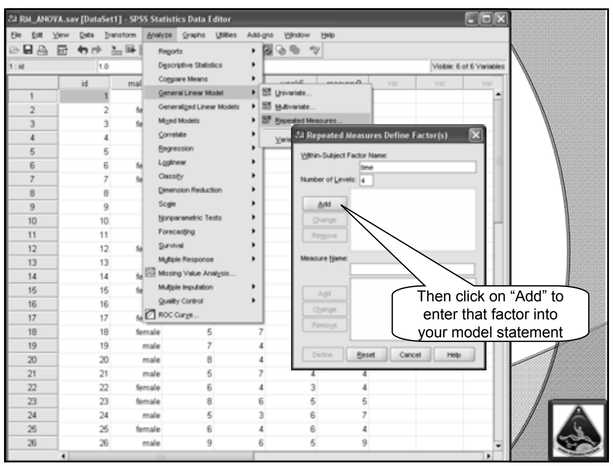
Using SPSS…

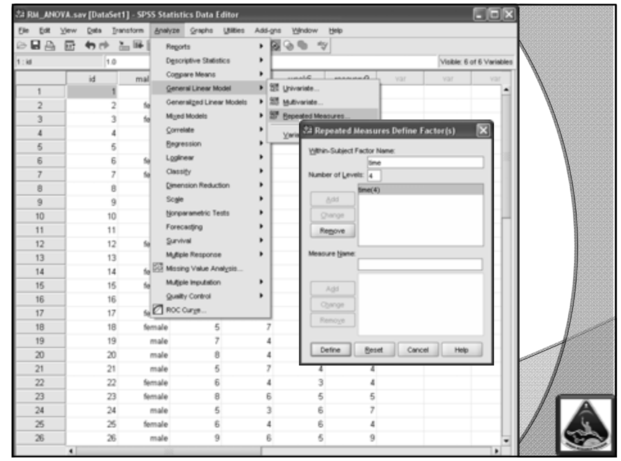


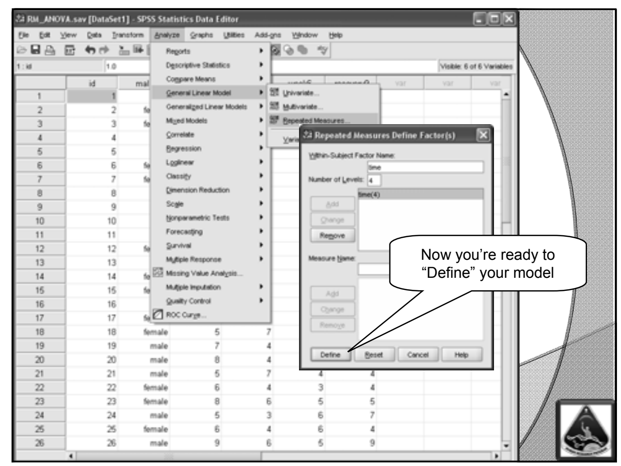
Subjects with ANY missing observation will be completely ignored in a purely repeated measures (fixed) ANOVA.

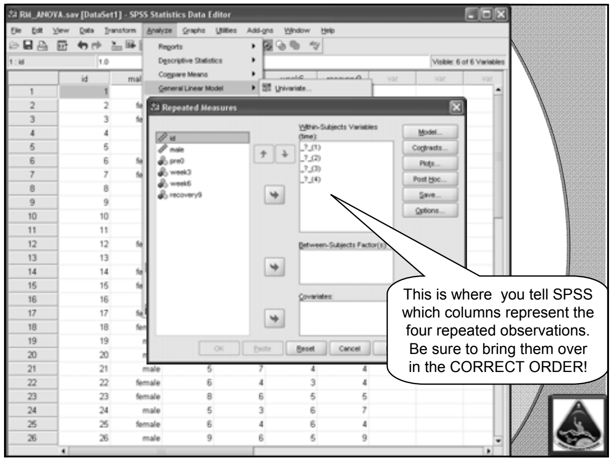
This can be a big problem with small n! We have n=34 here, but 4 are missing at least one observation… so Study n=30

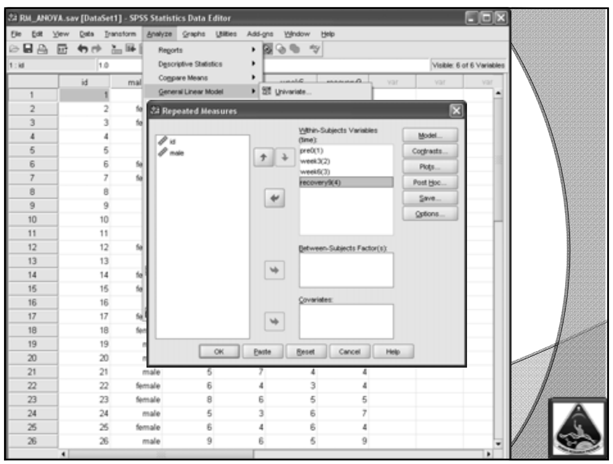

RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Repeated Measures Define Factor(s)
Give your factor a name that is meaningful to you
..and enter the number of repeated observations here

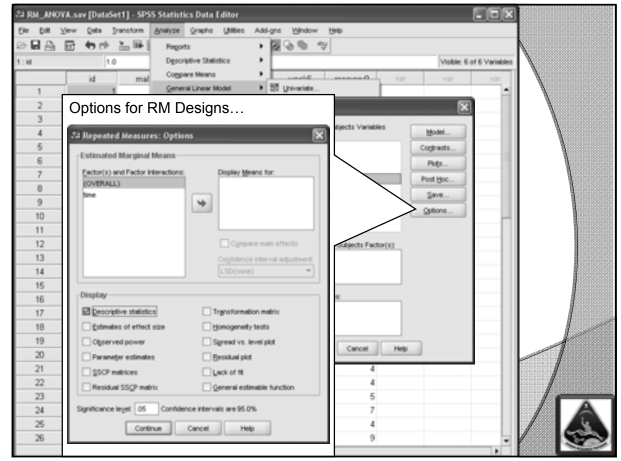

RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Repeated Measures Define Factor(s)
Then click on “Add” to enter that factor into your model statement

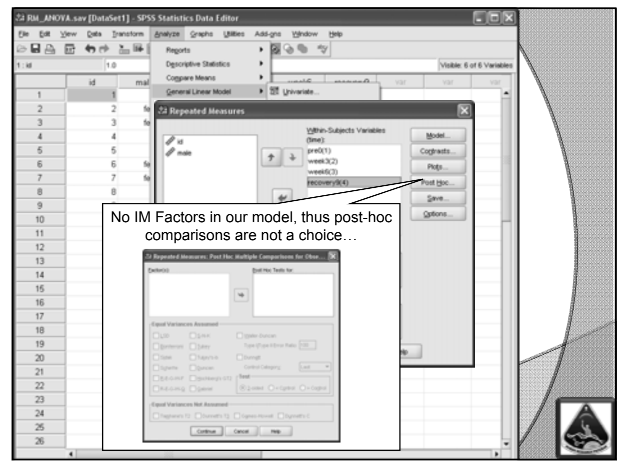

RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Repeated Measures Define Factor(s)

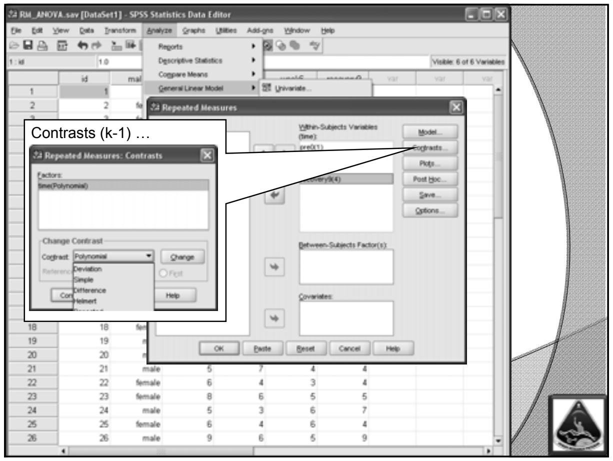

RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Repeated Measures Define Factor(s)
Now you’re ready to “Define” your model

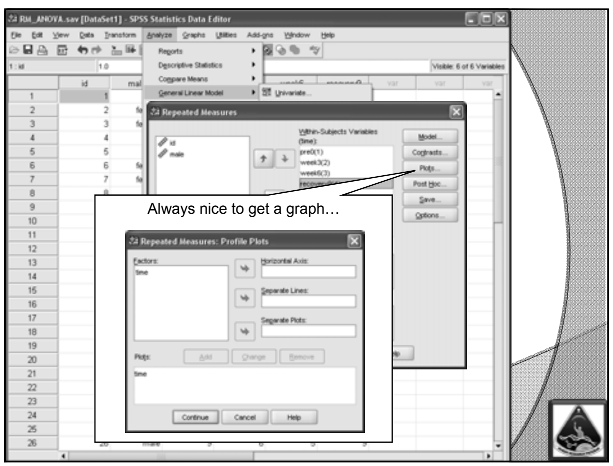

RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Repeated Measures
This is where you tell SPSS which columns represent the four repeated observations. Be sure to bring them over in the CORRECT ORDER!

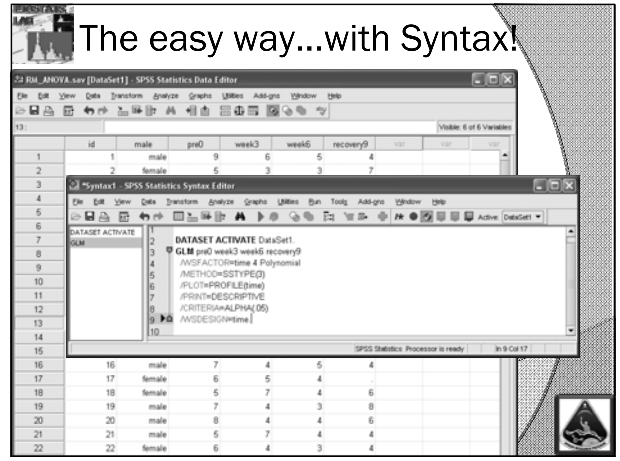

RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Repeated Measures

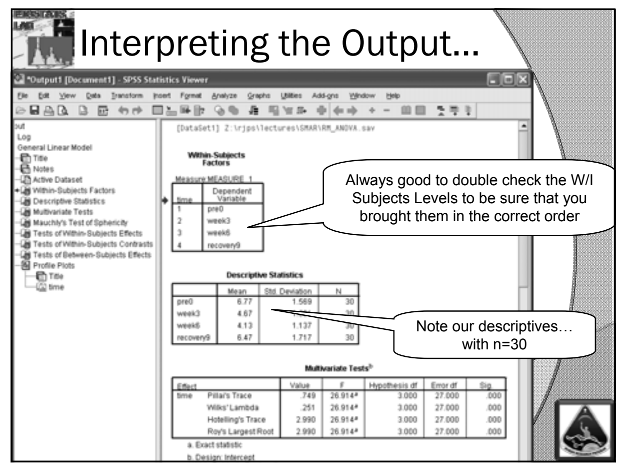

RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Options for RM Designs…
Repeated Measures: Options


RM_ANOVA.sav [DataSet1] - SPSS Statistics Data Editor
Repeated Measures
No IM Factors in our model, thus post-hoc comparisons are not a choice…

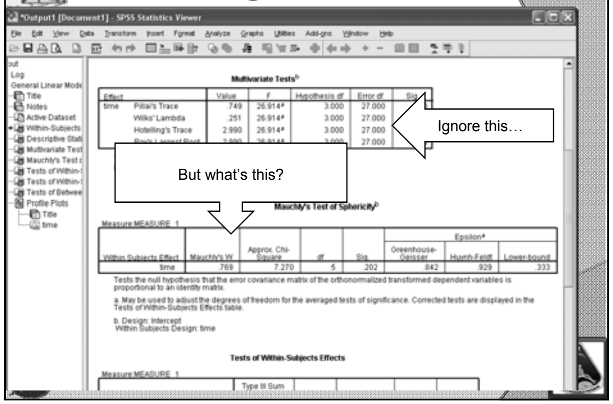

Contrasts (k-1) …
Repeated Measures: Contrasts
Change Contrast
Deviation
Simple
Difference
Helmert


Always nice to get a graph…
Repeated Measures: Profile Plots


The easy way…with Syntax!
DATASET ACTIVATE DataSet1.
GLM pre0 week3 week6 recovery9
/WSFACTOR=time 4 Polynomial
/METHOD=SSTYPE(3)
/PLOT=PROFILE(time)
/PRINT=DESCRIPTIVE
/CRITERIA=ALPHA(.05)
/WSDESIGN=time.


Interpreting the Output…
Always good to double check the W/I Subjects Levels to be sure that you brought them in the correct order
Note our descriptives… with n=30
Descriptive Statistics
Multivariate Tests

Interpreting the Output…



Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df
time	Pillai's Trace	.749	26.914a	3.000	27.000
	Wilks' Lambda	.251	26.914a	3.000	27.000
	Hotelling's Trace	2.990	26.914a	3.000	27.000
	Roy's Largest Root	2.990	26.914a	3.000	27.000

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

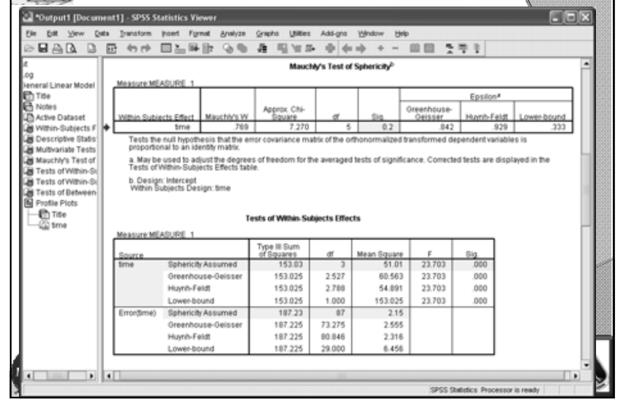
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a] Greenhouse-Geisser	Huynh-Feldt	Lower-bound
time	.769	7.270	5	.202	.842	.929	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept
Within Subjects Design: time

Interpreting the Output…



Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a] Greenhouse-Geisser	Huynh-Feldt	Lower-bound
time	.769	7.270	5	0.2	.842	.929	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

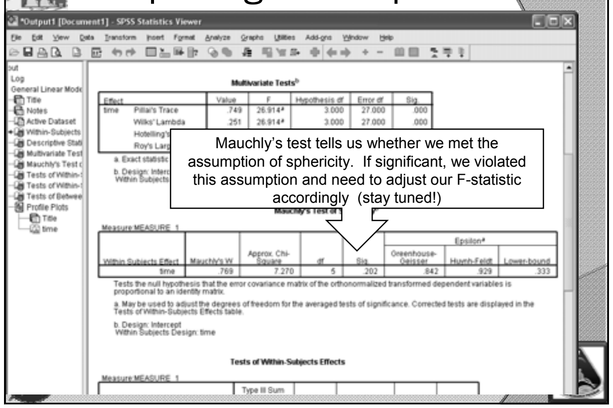
b. Design: Intercept
Within Subjects Design: time

Tests of Within-Subjects Effects

Measure:MEASURE_1

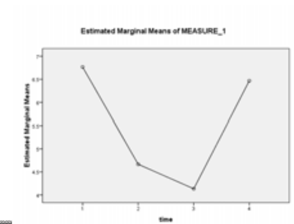
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
time	Sphericity Assumed	153.03	3	51.01	23.703	.000
	Greenhouse-Geisser	153.025	2.527	60.563	23.703	.000
	Huynh-Feldt	153.025	2.788	54.891	23.703	.000
	Lower-bound	153.025	1.000	153.025	23.703	.000
Error(time)	Sphericity Assumed	187.23	87	2.15		
	Greenhouse-Geisser	187.225	73.275	2.555		
	Huynh-Feldt	187.225	80.846	2.316		
	Lower-bound	187.225	29.000	6.456		

Interpreting the Output…



Let's look at our graph…

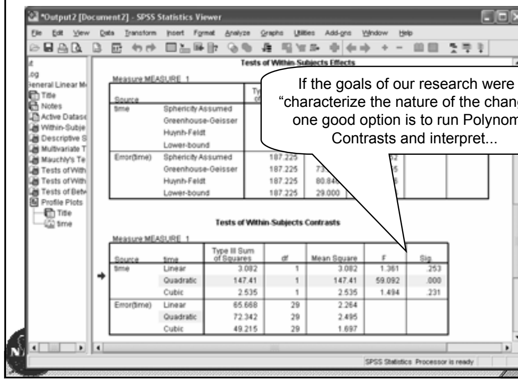
- Performance started high, dropped during bedrest, and seems to have returned by recovery.
- Different Options on how to follow-up with our significant ANOVA telling us that "things changed"



What next?

- Like with the Oneway IM-ANOVA, we probably had a more in-depth research question in mind, other than “did things change?” Did we want to:
  - “Characterize the nature of the change?”
  - “Compare “Pre-” levels to all “Post- levels” and report on our findings?”
  - “Compare everything to everything else and hope that something, *anything,* is significant so that we can get a paper outta this study???”

Back to our Output…



Tests of Within-Subjects Contrasts

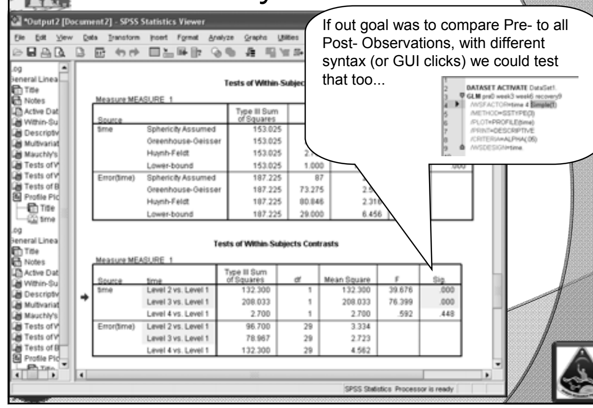
Measure:MEASURE_1

Source	time	Type III Sum of Squares	df	Mean Square	F	Sig.
time	Linear	3.082	1	3.082	1.361	.253
	Quadratic	147.41	1	147.41	59.092	.000
	Cubic	2.535	1	2.535	1.494	.231
Error(time)	Linear	65.668	29	2.264		
	Quadratic	72.342	29	2.495		
	Cubic	49.215	29	1.697		

What next?

- Like with the Oneway IM-ANOVA, we probably had a more in-depth research question in mind, other than “did things change?” Did we want to:
  - “Characterize the nature of the change?”
  - “Compare “Pre-” levels to all “Post- levels” and report on our findings?”
  - ~~“Compare everything to everything else and hope that something, *anything,* is significant so that we can get a paper outta this study???”~~

Alternatively…



Tests of Within-Subjects Contrasts

Measure:MEASURE_1

Source	time	Type III Sum of Squares	df	Mean Square	F	Sig.
time	Level 2 vs. Level 1	132.300	1	132.300	39.676	.000
	Level 3 vs. Level 1	208.033	1	208.033	76.399	.000
	Level 4 vs. Level 1	2.700	1	2.700	.592	.448
Error(time)	Level 2 vs. Level 1	96.700	29	3.334		
	Level 3 vs. Level 1	78.967	29	2.723		
	Level 4 vs. Level 1	132.300	29	4.562		

Contrasts with RM Factors

- Contrasts are powerful specific comparisons that can be run with Repeated Measures Factors
- They operate like "Post-Hoc" tests, but are called "Contrasts"
- With k levels of a Repeated Measures Factor, you can make k-1 contrast comparisons
  - So with 4 measures here, we can make 3 special contrast comparisons
- "Simple" and "Polynomial" are commonly used, but there are others too.

"Canned" Contrasts

- Polynomials – test for increasingly complex polynomial equations (linear, quadratic, cubic, etc.)
  - Useful to describe the trend, or nature of the changes
- Simple—compares all levels to a reference level
  - Common when there is a meaningful "pre" value
- Difference—compares each level (except the first) to the mean of all prior levels
- Helmert—compares each level (except the last) to the mean of all subsequent levels
- Repeated—compares each level (except the last) to the next subsequent level

Additional Contrasts

- Custom—users can also specify their own set of (k-1) contrasts per specific hypotheses
- Users can also perform Simple Effect Contrasts, like we demonstrated with IM-ANOVA, as long as an appropriate correction for the multiple comparisons are made…

```
GLM pre0 week3 week6 recovery9
  /WSFACTOR=time 4 Polynomial
  /METHOD=SSTYPE(3)
  /PLOT=PROFILE(time)
  /EMMEANS=TABLES(time) compare(time) adj(bonf)
  /PRINT=DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /WSDESIGN=time.
```

Next Time

- We'll discuss "doubly-repeated measures" designs, and run through an example or two.
- We'll run mixed-factorial designs, where we use a combination of RM and IM factors.
- We'll talk about including covariates in our models, and how that can be useful.

SMAR

HYPOTHESIS TESTING

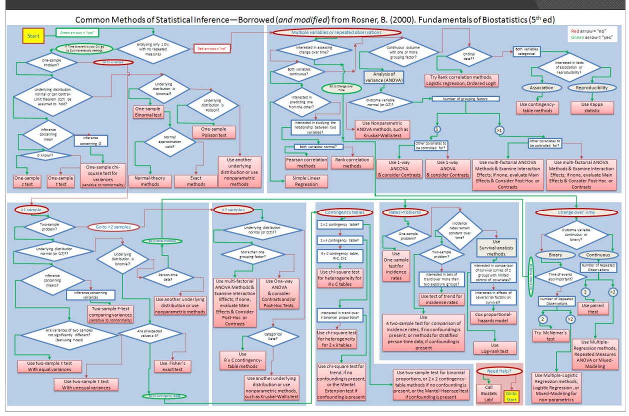
Truth Table

	The Truth is:	
	H_0 Really isTrue (there's no effect)	H_0 is Actually False (there is an effect)
You Rejected H_0 Due to a Statistically Significant Result	**Wrong Conclusion**	**Right Conclusion**
You Accepted H_0 Due to a Non-Significant Result	**Right Conclusion**	**Wrong Conclusion**

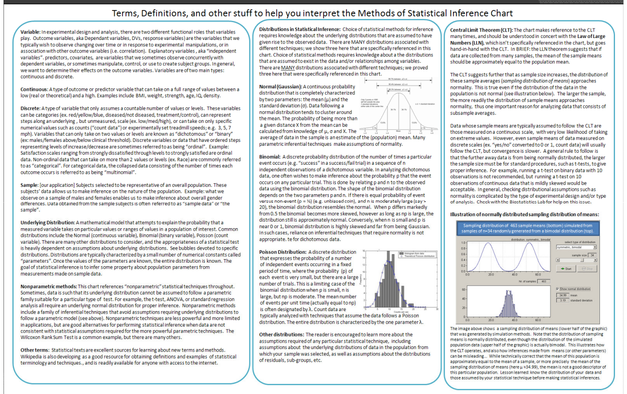
Alpha, Beta Type I & II Errors & Power

	The Truth is:	
	H_0 Really isTrue (there's no effect)	H_0 is Actually False (there is an effect)
You Rejected H_0 Due to a Statistically Significant Result	**Type I Error** Probability = α	**Power** Probability = (1-β)
You Accepted H_0 Due to a Non-Significant Result	Probability = 1- α	**Type II Error** Probability = β

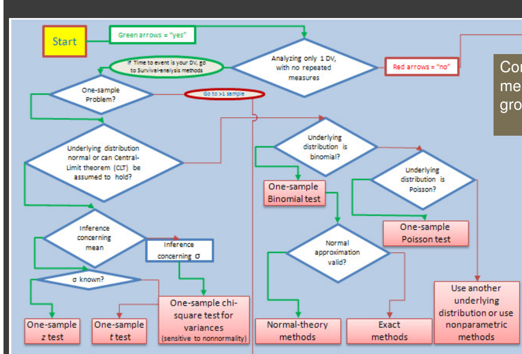
Common Methods of Statistical Inference



Terms, Definitions & Other Stuff

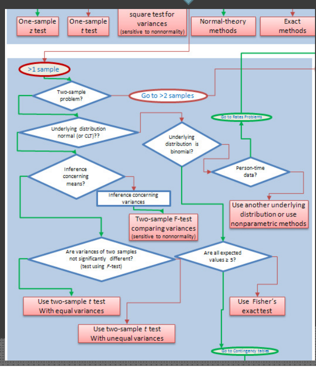


Where to begin (how to use the chart)?



Comparing means from 2 groups

Where to begin (how to use the chart)?

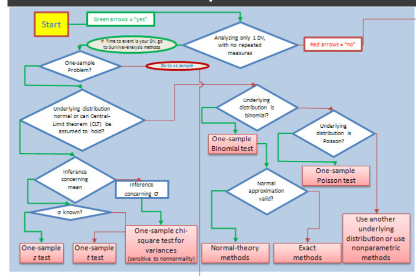


Comparing means from 2 groups(cont.)

Our data are scaled from a low number to a high one, and in the population, it tends to follow the bell-shape curve.

(What's missing from the flow chart?)

Another Example



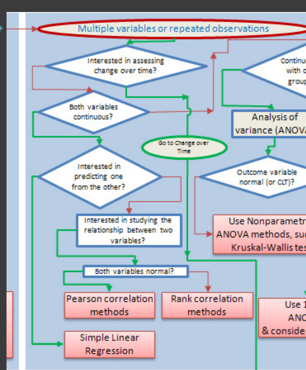
Want to know if BMD changes during bedrest are affected by an intervention, so we have 2 groups (Control, Intervention) and we measured their BMD Pre-bedrest, then again at 30, 60 and 90 days.

Want to know if the change over time is less (better) in the Intervention group relative to controls.

Another Example

Want to know if BMD changes during bedrest are affected by an intervention, so we have 2 groups (Control, Intervention) and we measured their BMD Pre-bedrest, then again at 30, 60 and 90 days.

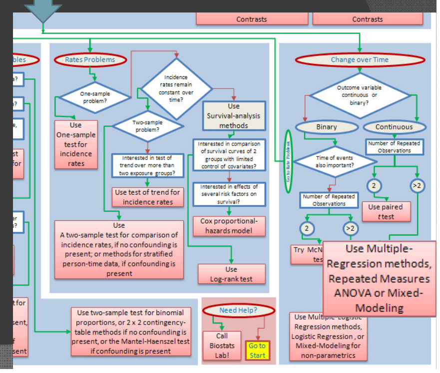
Want to know if the change over time is less (better) in the Intervention group relative to controls.



Another Example

Want to know if BMD changes during bedrest are affected by an intervention, so we have 2 groups (Control, Intervention) and we measured their BMD Pre-bedrest, then again at 30, 60 and 90 days.

Want to know if the change over time is less (better) in the Intervention group relative to controls.



Next Time

- Meet again at noon, Thursday, Sept. 24th
- Begin reviewing Hypothesis Testing using ANOVA, Regression, or Other topic per today
- PPT Slides & "Screenshots" from Statistical Software
  - Promise… no hand calculations & minimal formulae!
  - Promise… fun & applied, with enough "meat" to get you started and keep you statistically-safe
    - Or at least enough to know when it's time to call us!!